Understanding, Analysis, and Comparison of Convolutional Neural Network Architecture

Reid Chen, Yuehan Qin, Jijie Zhang

Introduction [Reid Chen, Yuehan Qin]

Neural networks, which are widely applied in deep learning, are inspired by neuron connectivity patterns that are similar to animal cortex structure. A specialized neural network, convolutional neural network (CNN), takes the advantage of ”convolution operation”. This operation, realized by a kernel or a filter, calculates the weighted sum of pixel information in a certain area defined by the size of the kernel. The kernel shifts across the image to retrieve information of the whole image. The benefit of using kernels, as used in CNN, instead of flattening the input image is that weights corresponding to a kernel can be shared by the whole image, meaning there are less parameters to be trained and hence improve the network performance.

Since the emergence of LeNet-5, convolutional neural networks have become an indispensable element of deep learning and computer vision, especially after the unprecedented results of AlexNet performed during the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Understanding these classic convolutional neural networks thus becomes an essential step of entering the academia of deep learning. We plan to compare and analyze the architectures of famous CNNs and try to understand the basis of why CNN is able to tackle image classification problems, even better than an average human.

Gradient-based Learning Applied to Document Recognition [Yuehan Qin]

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Link

LeCun et al. reviewed and compared several methods applied to handwritten character recognition. They found that CNNs, which are specifically designed for 2-dimensional shapes, outperform all other techniques. Specifically, CNNs are effective in making input patterns invariant with respect to transformations and distortions. In convolutional networks, shift invariance is automatically obtained by forcing the replication of weight configurations across space. Local receptive fields, shared weights, and spatial or temporal subsampling in CNNs are used to guarantee the invariance. Local receptive fields are able to extract elementary visual features such as oriented edges, endpoints, and corners; shared weights help apply elementary feature detectors that are useful on one part of the image to the entire image, and spatial subsampling reduces resolution of the feature map and reduces the sensitivity of the output to shifts and distortions.

Through review and comparison between different methods for document recognition, CNNs are found to be the most effective gradient-based learning technique for handwritten character recognition. CNNs are proved to be able to eliminate the needs for hand-crafted feature extractors. One limitation of the paper is that data generating models and the maximum likelihood principle are not used to justify most of the architectures and the training criteria

described in the paper. According to LeCun et al., such "hard to justify" principles would put strong constraints on the system architecture and worsen the performance. In General, the claim is convincing. The graphs that show the comparison results add credibility to LeCun et al.'s argument and the use of flow charts of the approaches used in this paper makes their result obtainable. Personally, I think the approach is computationally tractable for realistic problems as all their results are gained from a standard handwritten digit recognition task and the claims are accurate.

ImageNet Classification with Deep Convolutional Neural Networks [Yuehan Qin]

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. Link

Krizhevsky et al. trained a large, deep convolutional neural network (AlexNet) in the ImageNet-2010 contest into 1000 different classes and achieved considerably better results than previous state-of-the-art. The overall CNN architecture used contains five convolutional and three fully connected layers with weights. They used a rectified linear unit (ReLU) and cross-GPU parallelization scheme which improves the training speed by halving the number of the kernels (or neurons) in each GPU and limiting GPU communications only in certain layers. In order to reduce error, local normalization which creates competition for big response-normalized activities among neuron outputs computed using different kernels, and for which they do not subtract the mean, is utilized after applying the ReLU nonlinearity in certain layers. Overlapping pooling is used as well. Data augmentation using label-preserving transformation and dropout that consists of setting to zero the output of each hidden neuron with probability 0.5 are applied to avoid overfitting problem. Additionally, Krizhevsky et al. use weight decay of 0.0005 and follow the heuristic that divides the learning rate by 10 when the validation error rate stopped improving with the current learning rate as well as weight initialization from a zero-mean gaussian distribution with standard deviation 0.01.

Their result shows that large and deep CNN is capable of achieving record-breaking results on a highly challenging dataset using purely supervised learning. By adding new and unusual features to CNN, performance of it can be improved and faster training time can be achieved. Also, the significant problem for large and deep CNN, overfitting, can be prevented through data augmentation and dropout as well. Meanwhile, as Krizhevsky et al. state, producing labeled dataset that is large enough is important for making CNNs powerful and do what they could really do. This paper provides detailed methods for improving performance of large and deep CNN, which makes the approach computationally tractable for realistic problems. Accurate error rate for adding each feature is given and I think the claims are convincing and accurate. However, the number of very large labeled datasets is limited, which puts constraints on the application of the approach discussed in this paper as their result is gained through labeled dataset that is large enough to show the power of CNN. In the real world condition, it may be too expensive to label such a huge dataset unless a better and more efficient labelling approach is developed.

Visualizing and understanding convolutional networks [Reid Chen]

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. Link

In the article, Visualizing and Understanding Convolutional Networks, Zeiler & Fergus investigated the reason behind the outstanding performance of CNNs and explored how to improve the network proposed by Krizhevsky et al. To answer these questions and to understand CNN better, Zeiler & Fergus adopted a visualization tool called Deconvnet which is an inverse of CNN, meaning it can map features at a certain layer of the network to the image space, allowing us to see which part of the image contributes to the result at the layer the most. By examining each layer's visualization in the AlexNet and comparing these visualizations with the original input, Zeiler & Fergus showed that each convolutional layer could detect a certain pattern of the input image. These patterns are also composed together and become more and more stable as CNN goes deeper. Moreover, Zeiler & Fergus also applied an occlusion test, where a gray occluder masks a certain part of the input image. By examining a function between the mask's location and the output's accuracy, Zeiler & Fergus found that the accuracy drops significantly when the occluder masks the target object. Whereas when the occluder masks an irrelevant location, the accuracy is not influenced by much. Therefore, they conclude that CNN makes a prediction based on a local context instead of a global context. The visualization of each layer provides insight into how well the design of each convolutional layer. Zeiler & Fergus noticed the visualization of the first and second layer of AlexNet detecting unused features. With this insight, they changed the kernel size and stride and achieved better performance.

Through visualization, the paper reveals what the kernel's job is and what the kernels at different layers are doing. The visualization tool resolves CNN's black box, providing researchers more insights about it and providing machine learning engineers a new tool to inspect and modify CNN architecture. Zeiler & Fergus actually used the tool to modify AlexNet by changing the first layer's filter size from 11x11 to 7x7 and changing the stride size from 4 to 2. The data provided by them proved that changing the network according to the visualization results can, in fact, significantly outperform the performance of vanilla AlexNet. Moreover, the authors also provided evidence on the generalization of their CNN. The new architecture works well (near the state of the art performance) on their original ImageNet dataset and other datasets like Caltech-256. Their architecture is absolutely computationally tractable since they successfully trained it using a single GTX580 GPU, which is much slower than the current GTX 30 series GPU. As we know, training a deep neural network with large input takes a large amount of computational power and time. The finding of Zeiler & Fergus provides "more matter with less art" to the development of neural networks, saving developers and researchers many times from empirical approaches like training new networks with different designs.

Very Deep Convolutional Networks for Large-Scale Image Recognition [Reid Chen]

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Link

Like other classic networks, the network proposed by Simonyan & Zisserman (VGG-net) also participates in the ILSVRC competition. Therefore, it requires an input of 224x224. Different from many other networks, it mainly uses 3x3 filters. AlexNet, or the state-of-the-art architecture before Simonyan & Zisserman, used kernel size variates from 11x11 to 7x7 and had around 10 layers. Simonyan & Zisserman investigated the performance of deep convolutional neural networks (CNN) of 16-19 layers with small kernel size. They found that more convolutional layers with smaller kernels allow more non-linearity, which could improve the performance. This is because the ReLU activation function is added between every two

convolutional layers. They also noticed three convolutional layers with a 3x3 kernel could capture the same amount of information as a 7x7 kernel layer does with a stride of 2 (proposed by Zeiler & Fergus) while at the same time introducing fewer parameters to train. The VGG net's performance proves that the author's theory about smaller kernels and more layers is correct. The network achieved state-of-the-art accuracy and can be generalized to other datasets.

As mentioned by the authors, their paper is the first one that evaluates the performance of a small kernel on a large scale dataset. Moreover, it is also one of the first papers that increases the number of layers by a large step. Their theory about smaller kernel size is proved by their network's performance in the ILSVRC competition. The universal approximation theorem (1991) states that a network with three layers, enough nodes, and correct weights can approximate any function. This paper takes a different approach. Instead of adding more nodes into the first three layers, they choose to use fewer nodes (smaller kernel) and even more layers to approximate the true image classification function. Simonyan & Zisserman used this paper's result to show that the idea of going deep (shared by Szegedy et al.) is the next direction of improving CNN.

Going deeper with convolutions [Jijie Zhang]

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the *IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. Link

One of the main ways to improve the performance of deep neural networks is to increase their depth and width with large training data. However, it can cause overfitting and expensive computations. One idea is introducing sparsity such as replacing a fully connected layer with a sparse one. However, it does not reduce computations effectively. Due to the inefficiency of calculating non-uniform sparse data structures, the article proposes a new CNN structure codenamed Inception. Basically, this Inception contains a layer that stacks 1x1, 3x3, 5x5 convolutions with a 3x3 max-pooling layer. Using Inception modules, it applies dimension reductions and projections when there are too many computations. In theory, this sparse structure is supported by Hebbian theory which summarized as cells that fire together, wire together. This architecture has proven to be effective as the base network in localization and object detection. In addition, a network based on this structure, GoogLeNet, ranked the best in ILSVRC 2014 Classification Challenge and ILSVRC 2014 detection Challenge along with fewer computations. It is more precise and far deeper than past winning models like AlexNet but with fewer computations and parameters.

As Inception modules stack up to capture more high-dimensional features, the ratio of 3x3 and 5x5 convolutions will increase which eliminates sparsity and leads to computational blow up. In order to reduce cost and can keep sparse in most places, it applies a 1x1 convolution before 3x3 convolutions and after max pooling layer to lower dimensions. On the other hand, even the Inception architecture has proven to be successful in competition, it is still questionable about its reliance on guiding principles. Thus it requires more analysis and verifications in the future.

Deep residual learning for image recognition [Jijie Zhang]

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image

recog- nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. Link

Increasing layers in neural networks can improve performance is incorrect as it can cause degradation problems that more layers can produce more errors eventually. A deeper neural network is therefore more difficult to train. Deep plain nets are found to have low convergence rates, which affects the training error reductions. The article proposes a new neural network architecture to improve training called deep residual learning framework, ResNet. It introduces a residual block which does not learn weights after layers but only calculates the difference between input and output which is called skip connections. After applying ResNet architecture in testing ImageNet dataset, degradation problem is resolved as learning errors decrease with increasing layers. Also, as traditional CNN loses some data when transfering information, ResNet only learns input and output so that it prevents this problem. The calculation between input and output in the skip connection does not add additional parameters and computations while improving training speed and model precisions. In addition, the author mentions two-layer and three-layer ResNet. The later one applies the 1x1 convolution concept from Inception Net which reduces dimensions. Using a 152-layer ResNet, the network becomes the deepest among all models and won the best place in ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in ILSVRC & COCO 2015 competitions.

ResNet is phenomenal. Google combined this architecture with Inception network and proposed new models like Inception V4 and Inception-ResNet-V2. These two achieved incredible results with a 3.08% error rate in processing ILSVRC dataset. Currently, one potential problem is while using ResNet to explore over 1000 layers, the result is worse than the one of 110 layers. The author argues it is from overfitting and future work should research on whether stronger regularization can improve on.

Conclusion [Jijie Zhang]

The invention of convolutional neural networks and its varieties has continuously shown outstanding results in image recognition and computer vision tasks in the last few decades. From this point, those different networks have achieved better performance by variating its structures instead of blindly stacking up layers or increasing parameters. However, currently, the design of CNNs is mainly based on experience. This means that guiding principles or formulated methods of setting hyperparameters in a neural network still need to be discovered. A well-known example of CNN application is AlphaGo, which can beat Go grandmasters. The future goal will be finding better structures and proving architecture design guiding principles.