School of Informatics

Informatics Research Review
Practical Advances in Large-scale Image Classification on Deep Learning

[redacted]

January 2021

Abstract

The review paper provides a comprehensive overview of the notable practical improvements in large-scale image classification tasks using deep learning models. The aim is to provide the intended audience a way to track the developments of this field and encourage further algorithmic innovations. Starting with the pioneering work[1], the literature research is categorized into three groups according to the direction of improvements: network architecture, optimization techniques, and skip connections. The discussion and potential future works are presented in the conclusion section.

Date: Friday 22nd January, 2021

Supervisor: [redacted]

Contents

1 Introduction

The considerable variability of objects in the real world brings immense difficulty to visual recognition tasks. Image classification as a fundamental task in visual recognition categorizes images into one of several predefined classes[2], having been widely applied in many fields. With the developments of computer hardware[3] and large-scale dataset[4], the popularity of deep learning approaches continue to rise to accomplish the task. In particular, convolutional neural networks (CNNs) have led to a series of breakthroughs for large-scale image classification[5]. CNNs gain a compelling advantage in learning to solve complex image classification tasks as they have many more hidden layers to build powerful layers of abstraction. A large amount of evidence[6][7][8][9][5][3] since 2012 reveals that deep learning approaches are of great importance in the advances in image classification on the large-scale dataset and novel approaches are continuously proposed based on the previous works.

Driven by the importance of the topic and the absence of comprehensive review papers focused on large-scale image classification at the time of writing, this paper aims to provide a comprehensive synthesis of the representative research and critically analyses the key breakthroughs in the area of large-scale image classification with CNNs. The target audience is the researchers working on visual recognition as well as anyone interested in better understanding the evolution and the current state of this area. The key goals are to highlight the representative improvements through the development, provide the audience a way to track the progress of the state-of-the-art image classification, and inspire further algorithmic innovations in this area.

To this end, we will investigate the following questions:

1) What are the state-of-the-art deep learning approaches in large-scale image classification during the evolution process?

2) What are their practical improvements supported by experimental results?

3) Do these approaches have any limitations, and if so, which?

The paper will focus on the works that present the notable practical improvements supported by experimental results on the same benchmark large-scale dataset ImageNet[4]. We will not discuss those papers where their image classification experimentation was conducted on other datasets, such as MNIST, CIFAR, and so on, for the reason that these datasets are relatively small and the relative performance between these similar models cannot be compared and evaluated scientifically.

The paper is further divided into four sections and each section is organized as follows. Section 1 mainly introduces the research background, research problem, and objectives and scopes. Section 2 provides the brief background information that a reader need to understand the review. Section 3 first introduce the revolutionary work of Krizhevsky et al.[1], followed by the review of the symbolic improvements since then. According to the direction contributing to the improvements, the key papers are further categorized into three groups, namely network architecture, optimization techniques, and skip connections. Finally, Section 4 provides a summary of the research and discusses possible future extensions.

2 Background

2.1 Large-scale Image Classification

Image classification refers to the process of categorizing and labeling groups of pixels or vectors within an image into one of several predefined classes based on specific rules[2]. Datasets for the large-scale image classification task are large and complex. It typically contains hundreds of object categories and millions of images. In this paper, we investigate the works conducted on large-scale benchmark dataset ImageNet[4], the subset of which used for image classification consists of approximately 1.2 million training images, 50,000 validation images, and 150,000 testing images over 1,000 classes. Each image contains one ground-truth label representing its classes.

In the image classification task, algorithms may identify a list of object classes (up to 5) shown in each image as there could be multiple objects in the image. The performance of classification is evaluated based on the comparison of predicted classes and the ground truth label for the image. Mathematically, each image i has a single ground truth label C_i. The algorithms produce a list of labels c_{i1}, ... c_{i5}. If there is a j such that $c_{ij} = C_i$, the classification is considered to be correct. In practice, the metric top-5 error E is used to measure the error of the algorithm made in testing images. The error of a single prediction d_{ij} is assigned to 1 if $c_{ij} \neq \mathrm{C}_i$ for j from 1 to 5, and 0 otherwise. The error of an algorithm is the average number of misclassification of testing images:

$$E = \frac{1}{N} \sum_{i=1}^{N} d_{ij}$$

where N is the total number of testing images.

2.2 Convolutional Neural Networks

Concept

Convolutional neural networks (CNNs)[10] are a specialized kind of neural network for processing grid-like input data, such as time-series data, image data, and so on[11]. CNNs are inspired by the structure of the visual cortex in the brain[12] and designed to learn spatial hierarchies of

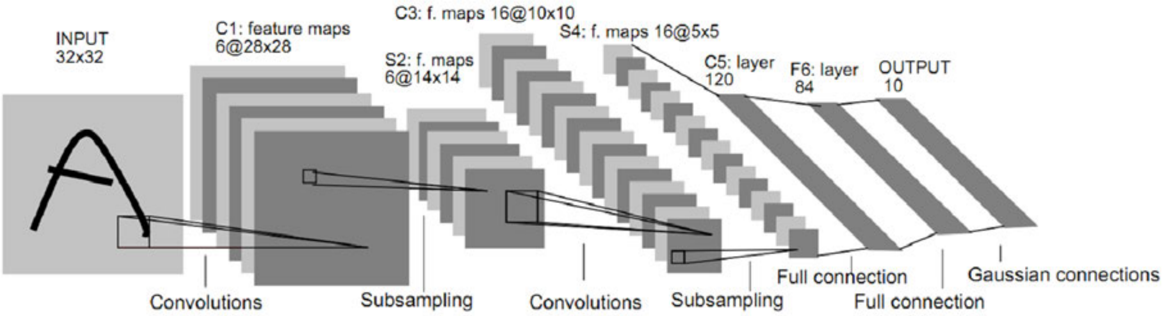

Figure 1: Example of Typical CNN Architecture[14].

features from low to high-level patterns in an automatic and adaptive manner[13]. They consist of neurons that have learnable parameters and each neuron receives some inputs, performs a dot product, and optionally follows it with non-linearity. In the image classification task, the network takes the raw image pixels as inputs and output the predicted classes and their scores.

Network Architecture

There are many variations of the network architecture of CNN. In general, it consists of three types of layers, namely convolutional layer, pooling layer, and fully connected layer. The convolutional and pooling layers are grouped into modules and perform feature extraction from inputs. There could be different numbers of modules stacked on top of each other and form a deep network. The fully connected layer typically follows the last module and maps the extracted features into the final output.

Figure 1 shows the typical network architecture of CNN for image classification. The input image is fed into several convolutional and pooling layers, where image features are extracted and subsampled to feature maps before sending into the later fully connected layers. Finally, the classification result is produced by the last fully connected layer. To improve the classification performance of CNNs or reduce the computational cost during training, several changes have been made to the network architecture, which is discussed in section 3.2.

3 Review of Main Developments

3.1 A Seed Paper

2012 witnessed the significant turning point for large-scale image classification, where large and deep convolutional neural networks entered the scene[1]. Although deep learning related algorithms have a long history since 1965[15], the research of neural network had developed slowly as it was considered that deep multi-layered networks were hard to train due to the lack of powerful hardware and optimization methods. It wasn't until 2012 that the pioneering work of Krizhevsky et al.[1] was proposed on using a large, deep CNN to classify the large-scale images in the ImageNet[4] in the context of deep learning, achieving groundbreaking results.

The authors briefly discussed the dramatic complexity of the visual recognition task, the attractive characteristic of CNNs contributing to the task, such as large learning capacity, few connections and parameters, and the advantages of GPUs in the training of large CNNs. More-over, they trained a large CNN with five convolutional and three fully-connected layers and employed regularization technique Dropout to reduce overfitting. The experimental results on ImageNet showed the model achieved by then the best results with top-5 test error rate of 15.3% with less training time, beating all traditional, non-deep learning models in image classification like[16][17] with at least 9% improvement. Their success was brought about mainly by the use of deep CNN with highly-optimization GPU implementation of 2D convolution. To reduce the large computational cost and the overfitting problem resulting from the number of layers, they carefully designed the network architecture by introducing novel or unusual features. Firstly, the network was trained on multiple GPUs, and Rectified Linear Units (ReLUs)[18] were in the place of tanh units after the neuron output to save the training time. Secondly, local response normalization and overlapping pooling were employed to improve the generalization ability and reduce error rates. Finally, data augmentation and Dropout effectively combat overfitting brought by 60 million model parameters.

The practical benefits of the groundbreaking work of Krizhevsky et al. have greatly motivated the majority of works discussed in this review and inspired the following innovations of deep learning-based methods in large-scale image classification.

3.2 Network Architecture

The work of Krizhevsky et al.[1] on deep CNN demonstrated state-of-the-art classification performance on the large-scale benchmark dataset, while there is no clear explanation and illustration of how CNNs work and why CNNs perform so well. In this way, their development and progress can only rely on constant trial-and-error. To address these issues, Zeiler and Fergus[6] introduced a new visualization technique using deconvolution[19] that gives insight into the hidden feature extraction layers of the network and the operation of the classifier. They noticed that the features in the intermediate layers show many informative characteristics of features such as compositionality, increasing invariance, and class discrimination as the layers ascend. Through the visualization, they debugged the potential problems in the network architectures of [1] - little coverage of the middle frequency information by the first layer filters and aliasing artifacts resulting from the large stride in the first convolutional layer. Besides, they found increasing the size of the middle convolutional layers could dramatically improve the performance of the experiments. From the findings, they finally proposed a new CNN model with 7×7 filters in the first layer and stride 2 in the first and second layers based on the network architecture of [1]. They increased the number of convolutional layers into 6, modified the number of neurons in the hidden layers, and decreased stride. The experimental results on ImageNet turned out that explored the modified network architectures significantly outperformed the single model of [1] by 1.7%, achieving the best-published performance with the top-5 error of 14.8%. However, the authors only visualized a single channel instead of the entire layer, which might cause the loss of important information in the layer level, affecting the final results.

Based on Hebbian principles, Szegedy, Liu, et al.[7] proposed an efficient deep CNN architecture named GoogLeNet for large-scale image classification and made significant progress by halving the classification error compared to [6]. As proved in the work of Zeiler and Fergus [6], increasing the size of deep neural networks can significantly improve their performance. However, they bear the cost of overfitting caused by a larger number of parameters and the dramatic increases in computational resources. To solve these issues, they were inspired by the theoretical work by Arora et al.[20] and changed the fully connected convolutional architecture into a sparsely connected one, allowing for increasing the size of the network while staying constant computational cost. Specifically, they employed additional 1x1 convolutions as dimension reduction blocks before each computationally costly 3x3 and 5x5 convolutions and after max-pooling layers in the network to reduce the parameters together with rectified linear activation. Besides, they also solved the common unstable gradients problem in deep CNNs by adding auxiliary classifiers connected to intermediate layers, increasing the gradient signal that gets propagated back and providing additional regularization. With these inception modules, they increased the number of layers into 22 and achieved 6.67% top-5 error rate on ImageNet classification with less computational cost, which is a 56.5% relative reduction compared to [1], and about 55% relative reduction compared to the best approach[6] in the previous year.

With its excellent performance, GoogLeNet has been studied and used by many researchers. Inspired by the performance of residual network[5] discussed in section 3.4, the improved inception architecture of GoogleNet[21], where the original 5x5 convolutions were replaced by computationally cheaper 3x3 convolutions, was further combined with skip connections. The experimental results of [8] on ImageNet classification showed the hybrid residual inception model accomplished the improved results with top-5 error rate of 3.08%, while the high computational cost recurred due to the ensemble of networks.

Recent remarkable development on network architecture was made by Hu et al.[9]. Unlike previous works that treated feature channels as of the same importance, they focused on the relationship of feature channels and proposed a novel network architecture called SENet, consisting of stacked"Squeeze-and-Excitation" (SE) blocks. The main idea was to obtain the importance of each feature channel through learning and then emphasize the useful features and suppressed trivial features for the tasks according to the importance. In practice, the authors construct a SE block to perform recalibration on the feature maps obtained after a series of convolutions. Firstly, the feature maps were shrunk through spatial 2D dimensions by global average pooling to obtain the global distribution of channel-wise responses, allowing lower layers of the network to be able to use the information from the global receptive field. Then, the excitation operation employed a gating mechanism where the parameters were learned to explicitly model channel association and produce channel-wise weights representing the importance of each channel. Finally, the original feature maps were reweighted channel by channel and fed into subsequent layers. Classification results showed SENet generalized well to the large scale dataset ImageNet and reduced the top-5 error to 2.251% at slight additional computational cost, surpassing the previous leading model[21] by a relative improvement of 25%. Besides, the SE module has high applicability. It can be embedded in the building block unit of almost all models like GoogLeNet discussed previously, ResNet discussed in section 3.4, to obtain different types of SENet. While, the reason why the network with the SE module converges faster than the original network is not well explained, lacking clear justifications.

3.3 Optimization Techniques

The training of deep CNNs is more different than that of shallower CNNs as the depth could hinder the convergence of the network and cause several optimization problems. In this section, two important optimization techniques are discussed in terms of how they help improve the classification performance.

Batch Normalization(BN)[22] was proposed by Ioffe and Szegedy to address the amplified change in the distribution of network activations due to small changes of parameters in the training of deep neural networks. The optimization problem is referred to internal covariate shift problem. The idea of BN is to incorporate normalization in the model architecture, perform the normalization for training mini-batches and propagate the gradients through normalization parameters. It transformed the mean of each dimension of input data to zero and the variance to

1 so as to avoid data shift, meanwhile, scale and shift each normalized activation with learnable scale parameters α and β to ensure identity transform. Empirical results on ImageNet dataset showed that the batch-normalized deep CNN [8][5] achieved lower top-5 error with fewer training steps compared to the original networks mentioned and outperforms human raters. The main advantages of this method are 1) apart from addressing the internal covariate shift problem, it enables a higher learning rate and helps tackle the gradient vanishing problem by stabilizing the parameter growth 2) it also regularizes the model and reduces overfitting. While the limitation of this method is that it doesn't work well for the cases with a small batch size or binary classification with unbalanced classes.

The initialization of the network parameters is critical as it could have a negative impact on network convergence. A considerable amount of initialization schemes has been conducted to address the problem[5][23][24], most notably layer-sequential unit-variance (LSUV) initialization [24]. The idea of the LSUV initialization scheme is firstly pre-initialize weights of each convolutional layer and inner-product layer with orthonormal matrics and then normalize the weights to make the output variance of these layers to be one. It can be understood as an orthonormal initialization coupled with BN[22] applied on the first mini-batch. The unit variance normalization process is similar to BN, while the orthonormal initialization breaks the correlation between layer activations and was proved to be more efficient compared to the full BN. Besides, the experimental results on ImageNet also showed that this scheme enhanced the training of deep CNNs with less computational cost and the LSUV-initialized GoogLeNet learned much faster and produced the lower testing error all the time.

3.4 Skip Connections

Deep CNNs gain a compelling advantage in learning to solve complex visual tasks as they have many more hidden layers to build powerful layers of abstraction. Research[25] suggests that deep network architectures are intrinsically more powerful than shallow ones in certain problems at the theoretical level. Although increasing the depth of CNN could improve the performance in general, it potentially brings some problems as discussed in section 3.2, such as overfitting, vanishing gradients, high computational cost, and so on. Especially, the problem of vanishing gradients, and more generally, unstable gradients, are common during training, which hinders the learning and model convergence of the model from the beginning, resulting in dramatic performance degradation. Driven by the importance of depth, a large amount of researches has been conducted to address the problem and subsequently improve the performance, most notably introducing skip connections into the network.

The degradation problem of accuracy caused by the increased depth of CNN also has been studied and addressed in [5]. The authors proposed a deep residual learning network(ResNet), where a residual mapping $\mathcal{F}(x) + x$ is applied in the feedforward neural networks by inserting a shortcut connection with the earlier layers and later layers of the basic block and applying element-addition to the output y of deeper layer with the identity of the input of shallow layer x. Formally, the convolutional processing block with skip connection can be defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{1}$$

where $\{W_i\}$ denotes the set of weights and transformations until layer i, and $\mathcal{F}(\mathbf{x}, \{W_i\})$ indicates the residual mapping the stacked layers to fit. The identity mapping can be applied directly when the dimension of the output of residual function $\mathcal{F}$ are equal to that of identity x. Additional linear projection W_s should be conducted on the identity when the dimensions are inconsistent. Ensembles of many shortcut connections with the identity mapping compensate 1 in the backward propagation, ensuring that the gradient won't exponentially decrease to zero to avoid the gradient vanishes in deep models. Empirical results on ImageNet shows that the residual networks with increased depth are easier to train and converge to a relatively record-breaking accuracy with 3.57% top-5 error. However, a limitation of deep residual networks is the great amount of redundancy of layers as suggested in stochastic depth[26] the training can be improved by randomly dropping layers during training.

To improve the deep residual networks, the original authors further analyzed the effectiveness of ResNet theoretically and proposed a new residual unit[27] that creates a direct path through the entire network in both forward and backward information propagation rather than only within a single residual unit[5]. The idea was implemented by removing the ReLu activation in the original residual mapping and employing the same identity mapping as skip connections with the after-addition activation to make the information directly flow from one residual block into the next residual block. They empirically proved that the new unit made the training of 1000 layers easier by the quick reduction in training loss compared to the original ResNet, and achieve 1.1% lower error than the baseline ResNet-200.

Furthermore, inspired by the redundancy in the residual network, dense convolutional network (DenseNet)[3] was proposed to enhance the propagation and reuse of features, leading to the substantial reduction of the number of parameters. Instead of summing the identity function and the output of later layers as in ResNet[5], DenseNet leverages direct connections between layers to concatenate features from different layers, leading to better information flow. It breaks L layers and L connections in a traditional convolutional network and performs a dense connectivity pattern with $\frac{L(L+1)}{2}$ connections where each layer is directly connected to the proceeding layers that have the same feature-map size and the features from proceeding layers are concatenated as input to the subsequent layer. Mathematically, the input of the l-th layer is represented as:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{2}$$

where $[x_0, x_1, ..., x_{l-1}]$ denotes the concatenation of the feature maps produced in layers 0, ..., l-1, and H_l represents the composite function of BN[22], a rectified linear unit[18] and a 3x3 convolution. In this way, each layer can have direct access to the gradients, leading to improved information and gradient flow throughout the deep network. The experiments showed that DenseNet only required around $\frac{1}{3}$ of the parameters of ResNets[5] to obtain the same level of performance with less computational cost. While, DenseNet could be memory inefficient with the naive implementation of the architecture.

4 Summary & Conclusion

This review provide a comprehensive overview of practical advances in large-scale image classification on deep learning. It categorizes the progression into the turning point and the following main developments. The pioneering work of [1] that brought deep convolutional neural networks into the scene, has revolutionized image classification and motivated the continuous improvements:

· Modifications of the CNN architecture[6][7][8][9], especially on the convolutional layers, have been proposed that allow for building powerful layers of abstraction, resulting in the state-of-the-art classification performance.

· The training process of deep CNNs have been improved by some notable optimization techniques[22][24], solving the optimization problems and bringing a considerable error reduction.

- The leading performance can be achieved probably with less computational cost by employing skip connections in CNNs[5][27][3] to enhance the model convergence.

The practical advances in ImageNet classification are quantified by the evaluation measure top-5 error rate, and the error was significantly reduced from 15.3% by the pioneering work[1] to 2.251% by SENet[9], surpassing that of human raters.

The significant improvements were achieved based on the understanding of CNN architecture and further addressing the problems encountered in the training process. Driven by the high abstraction power of convolutional layers, a series of works focused on them. The work in [6] increased the size of the intermediate convolutional layers based on the visualization of feature maps, resulting in slight performance improvement but additional computational cost. The sparsely connected design of convolutional architecture, GoogLeNet and the improved model[7][8], alleviated computational burden, enabling a large increase of network size to improve classification accuracy. Different from the previous works[6][7][8] treating each feature channel as of the same importance, SENet[9] recalibrates the feature maps obtained from convolutional layers to emphasize the important features, accomplishing the leading performance.

Optimization techniques[22][24] were designed to solve the potential internal covariate shift problem in deep CNN. By normalizing the input mini-batches and introducing the learnable normalization parameters, BN[22] enables a higher learning rate and regularizes the model when incorporated into the mentioned models[8][5]. However, the additional parameters lead to increased computational cost, motivating the improved techniques[24]. It employed the same unit variance normalization as BN, but broke the correlation between layer activations by the orthonormal weight initialization, effectively improve the computation efficiency. However, the results of the weight initialization scheme are unstable on the large-scale dataset. Therefore, BN is more commonly used to solve optimization problem in CNN. It is highly expected that solutions can be proposed to reduce the computational burden resulted from extra parameter requirements of using BN.

Driven by the increased improvements brought from the depth of network layers, CNN tends to be deeper and deeper. To solve the degradation challenge resulting from the depth, several types of skip connections were proposed to enhance the information flow through the network and promote the model convergence. ResNet[5][27] leveraged identity mapping from the earlier layers to later layers, easing the training of the deep model. While the proven layer redundancy problem and decreased performance with Dropout regularization technique encourage further work on ResNet. Instead of summing the identity function and the output of later layers as in ResNet[5], DenseNet created direct connections across layers to concatenate features, enhancing the information propagation while requiring less computation.

Although the experimental results of these models are encouraging, almost none of the papers provide the theoretical basis for their success on large-scale image classification. It brought limitations to the model robustness and adaptability in different classification tasks. The slight modification of the complex model architecture could lead to a dramatic drop in the performance. Besides, there always exists a trade-off between the model performance and the computational cost, restricting the overall performance. Future works could focus on the theoretical analysis of the empirical advances accomplished by these convolutional-related modifications and shortcut connections, the new solutions to further address the problems of current models, such as alleviating the computation burden, promoting the effectiveness of BN combined with regularization techniques and so on, and the innovations of the training process.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[2] M Shanmukhi, K Lakshmi Durga, Madela Mounika, and Kurakula Keerthana. Convolutional neural network for supervised image classification. *International Journal of Pure and Applied Mathematics*, 119(14):77–83, 2018.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*. Citeseer, 2009.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *In Computer Vision–ECCV 2014*. Citeseer, 2014.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141. IEEE, 2018.

[10] Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[13] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

[14] 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.

[15] AG Ivakhnenko and VG Lapa. Cybernetic predicting devices. Technical report, PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGINEERING, 1966.

[16] Jorge Sánchez and Florent Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR 2011*, pages 1665–1672. IEEE, 2011.

[17] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010.

[18] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[19] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pages 2018–2025. IEEE, 2011.

[20] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pages 584–592, 2014.

[21] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[23] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[24] Dmytro Mishkin and Jiri Matas. All you need is a good init. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–13, 2016.

[25] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

[26] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in neural information processing systems*, pages 550–558, 2016.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.