Deep Learning Approaches to Image Classification

Samuel Olowofila
Department of Computer Science
University of Colorado Colorado Springs
CO, 80918 USA

Abstract

This paper delves into image classification, a pivotal area in computer vision, through supervised learning techniques using predefined category labels. Leveraging Artificial Neural Networks, particularly architectures like Convolutional Neural Network (CNN), Residual Network (ResNet), and VGG16, significant advancements have been made since the early 2010s. The study employs popular datasets such as MNIST, CIFAR-10, ImageNet, and Places. We introduce a basic CNN model and its deeper variant, and explore dual implementations of ResNet50 and VGG16, both in their original form and pre-trained on the Imagenette dataset, a subset of ImageNet. The results highlight the importance of model depth, skip connections, and the benefits of transfer learning. Overall, the research underscores the need for choosing architectures based on dataset and task specificities, reflecting on the complexities and advancements in image classification through deep learning.

Introduction

Image classification is a fundamental challenge in the computer vision sphere. It is a supervised learning challenge, charging researchers with the task of assigning labels to an image within the confines of a predefined set of categories. The importance of image classification has grown drastically owing to its widespread application in medical imaging, security and surveillance, retail, and the automotive industries, to mention a few. Dating from the early 2010s, there have been significant breakthroughs in this research area courtesy of Artificial Neural Networks(Lee et al. 2015). The results documented by investigators of neural networks in this research space have recorded varying margins of success with the Convolutional Neural Network (CNN)(He et al. 2016), Residual Network (ResNet)(Jian et al. 2016), VGG16(Liu et al. 2017), amongst others. Due to the large dataset required to train a deep neural network, the MNIST (Yann 1998), CIFAR-10(Krizhevsky, Nair, and Hinton), ImageNet(Deng et al. 2009), AND Places(Zhou et al. 2017) datasets have been the most widely applied datasets in this research domain. Here, different configurations of convolutional neural networks using the aforementioned datasets are examined.

Approach

In this research work, a baseline CNN model based on a simple convolutional architecture is developed alongside a variant that introduces more depth to the network for a more robust comparative analysis. To further the investigation, a dual-way implementation of the ResNet50 and VGG16 models were also developed, with each having a vanilla baseline version and a second variant pre-trained on the Imagenette(Howard and others 2020) dataset and fine-tuned using the Mnist and CIFAR-10 datasets. The Imagenette dataset is a subset of the larger ImageNet dataset that comprises ten easily distinguishable classes from the original ImageNet dataset representing the images in 160px and 320px resolutions. This investigation, however, adopts the 320px resolution option. Being a large-scale dataset, and due to resource constraints, the ImageNet dataset is not favorable for this experiment. Hence, the adoption of the Imagenette alternative.

CNN

The architecture of the baseline CNN model I developed is made up of two convolutional layers. The first layer implements 32 filters of size 3 x 3 with a padding value of 1, while the second layer implements 64 filters and retains the original filter size. Each of the convolution layers concludes with a max pooling operation using a 2 x 2 window dimension. This architecture also features two fully connected (FC) layers while introducing a 50% dropout after the first layer to help prevent overfitting. The FC layers comprise a dense later of 128 and 10 neurons, respectively. The ten neurons of the latter layer correspond to the 10 digital class labels of the datasets. Cross-entropy loss is the choice of loss function for this model since it is typical for classification tasks(Hui and Belkin 2020). DeeperCNN is the second convolutional model variant examined in this experiment. It features three convolutional layers, with the third layer exhibiting 128 feature maps, and adds an additional pooling operation to the network. DeeperCNN also integrates batch normalization after each convolutional layer while retaining two FC layers, with the first FC layer featuring 256 neurons. Furthermore, a second dropout layer is incorporated at a 30% rate after the output layer.

	Baseline CNN	**DeeperCNN**	**ResNet50**	**VGG16**
Input	1x32x32 (MNIST Resized)	1x32x32 (MNIST Resized)	3x224x224	3x224x224
Convolutional	32 filters, 3x3, padding=1	32 filters, 3x3, padding=1	Initial conv, 7x7, stride=2	Multiple 3x3 convs
BatchNorm	-	Yes (32 feature maps)	Yes	-
Activation	ReLU	ReLU	ReLU	ReLU
Pooling	Max pooling, 2x2	Max pooling, 2x2	Max pooling, 3x3, stride=2	Max pooling, 2x2
Conv Blocks	64 filters, 3x3, padding=1	64 filters, 3x3, padding=1	Multiple residual blocks	Multiple 3x3 convs
BatchNorm	-	Yes (64 feature maps)	Yes	-
Conv Blocks	-	128 filters, 3x3, padding=1	Multiple residual blocks	Multiple 3x3 convs
BatchNorm	-	Yes (128 feature maps)	Yes	-
Pooling	-	Max pooling, 2x2	-	Multiple
Fully Connected	64x8x8 to 128 neurons	128x4x4 to 256 neurons	Global avg pooling	3 Fully connected layers
Dropout	0.5	0.5	-	In final FC layers
Output Layer	128 to 10 neurons	256 to 10 neurons	Dense, softmax	Dense, softmax
Additional Dropout	-	0.3	-	-

Table 1: Architectural comparison of the models

ResNet50

The ResNet50(Jian et al. 2016) was introduced in a CVPR conference to address the vanishing gradient challenge typical of deep neural networks. This model introduces skip or residual connections whose essence is to bypass one or more layers in the network. The out-of-box model, as provided by Pytorch torchvision, was trained using the MNIST, CIFAR-10, and ImageNet Datasets. Furthermore, we implemented a modification to this model using the Transfer Learning concept(Hussain, Bird, and Faria 2019). This modified variant has the ResNet50 model stripped of its top (classifier) layers, implements Group Average Pooling (GAP) on the output of the network, and integrates a dense FC layer with 1024 neurons before implementing a ReLU activation function.

VGG16

This model was introduced by an Oxford research group named “Visual Geometry Group” (Simonyan and Zisserman 2014). VGG16 connotes the architecture of the model, comprising 13 convolutional layers and 3 FC layers. We train this vanilla model without pre-training on the MNIST, CIFAR-10, and Imagenet datasets, using the Sparse Categorical Crossentropy with logits as its loss function. Likewise, an adjusted variant of VGG16 was developed by first pre-training it on the Imagenette dataset, and detaching the prediction layer at the top of the network, flattening the output of the convolutional layers, and decking with two dense FC layers having 4096 nodes each, and a prediction layer with ten nodes corresponding to the class labels. The softmax activation function is then implemented.

Results

Throughout this experiment, the MNIST, CIFAR-10, and Imagenette datasets were kept constant to ensure a fair comparative analysis. Also, the loss function of choice for the entire work was Sparse Categorical Crossentropy. However, the input datasets were adjusted to fit the default input dimension required by the vanilla architectures of all models considered; 224 x 224 with 3 RGB channels, and normalized to the factor of 225 for the VGG16 and ResNet50, while a the datasets were adjusted to 28 x 28 with 1 channel for the Baseline CNN and DeeperCNN. The trainings were limited to 10 training loops or epochs due to computational cost and resource constraints. An excerpt of the results is recorded comparatively in Table 2.

Model	**Accuracy (%)**
Baseline CNN	99.11
DeeperCNN	79.02
Baseline VGG16	99.21
Baseline ResNet50	99.47
Pre-trained VGG16	99.91
Pre-trained ResNet50	99.84

Table 2: Performance comparison of the various models on MNIST datatset.

State-of-the-art

The fatality of the consequences of erroneous image classification in application areas such as self-driving autonomous cars(Ajenaghughrure, da Costa Sousa, and Lamas 2020) keeps driving the need for improvement in this research area(Nascimento et al. 2019). Hence, the continuous efforts by researchers towards the amelioration of existing outputs. Shifting from the traditional CNNs, the Vision Transformer (ViT) have proven to challenge the status quo in image classification tasks. It adopts transformer mechanisms which are typical for natural language processing exercises.(Dosovitskiy et al. 2020)

Conclusion

In the comprehensive exploration of various deep learning architectures for image classification, distinct differences in model structures and their implications were evident. The foundational Baseline CNN and its extended counterpart, DeeperCNN, demonstrated the core tenets of convolutional processing, showing how depth and added components like batch normalization can influence performance. In contrast, the advanced architectures of ResNet50 and VGG16, both renowned for their capacity to handle complex patterns in large-scale datasets, were pivotal in emphasizing the importance of depth and skip connections. The transfer learning experiments, especially with ResNet50 pre-trained on the Imagenette dataset, accentuated the power of leveraging prelearned features for new tasks, thereby potentially speeding up training and enhancing generalization. While each model had its unique merits, the overarching conclusion underscores the significance of choosing an architecture tailored to the specificities of the dataset and task at hand. The intricacies of these models, combined with the versatility of the datasets used, shed light on the multifaceted nature of image classification and the ever-evolving landscape of deep learning solutions."

References

Ajenaghughrure, I. B.; da Costa Sousa, S. C.; and Lamas, D. 2020. Risk and trust in artificial intelligence technologies: A case study of autonomous vehicles. In *2020 13th International Conference on Human System Interaction (HSI)*, 118–123. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard, J., et al. 2020. Imagenette. *URL https://github.com/fastai/imagenette* 5:11.

Hui, L., and Belkin, M. 2020. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*.

Hussain, M.; Bird, J. J.; and Faria, D. R. 2019. A study on cnn transfer learning for image classification. In *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*, 191–202. Springer.

Jian, S.; Kaiming, H.; Shaoqing, R.; and Xiangyu, Z. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, 770–778.

Krizhevsky, A.; Nair, V.; and Hinton, G. Cifar-10 (canadian institute for advanced research).

Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2015. Deeply-supervised nets. In *Artificial intelligence and statistics*, 562–570. Pmlr.

Liu, B.; Zhang, X.; Gao, Z.; and Chen, L. 2017. Weld defect images classification with vgg16-based neural network. In *International forum on digital TV and wireless multimedia communications*, 215–223. Springer.

Nascimento, A. M.; Vismari, L. F.; Molina, C. B. S. T.; Cugnasca, P. S.; Camargo, J. B.; de Almeida, J. R.; Inam, R.; Fersman, E.; Marquezini, M. V.; and Hata, A. Y. 2019. A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Transactions on Intelligent Transportation Systems* 21(12):4928–4946.

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

Yann, L. 1998. The mnist database of handwritten digits. *R*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40(6):1452–1464.

Supplementary Material

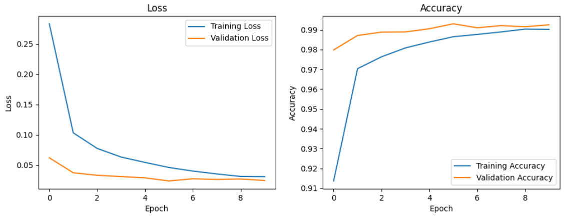

Figure 1: Baseline CNN with MNIST

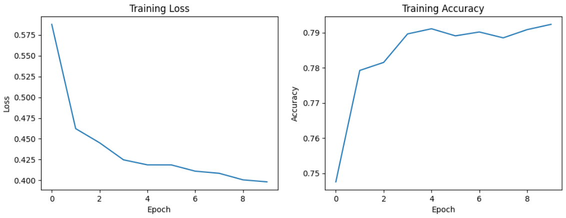

Figure 2: DeeperCNN with MNIST

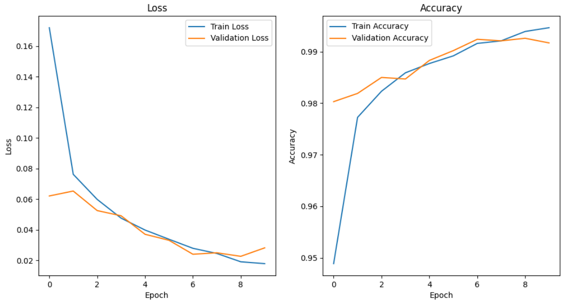

Figure 3: Baseline ResNet50 with MNIST

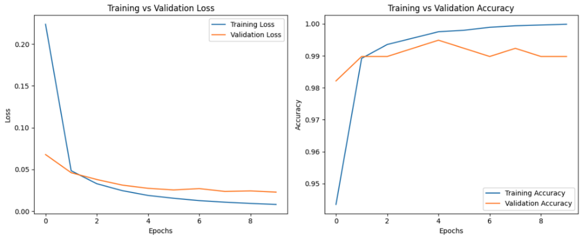

Figure 4: Pre-trained and fine-tuned VGG

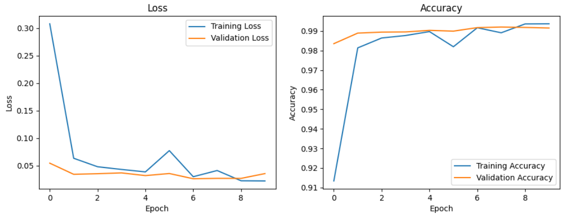

Figure 5: Baseline VGG with MNIST

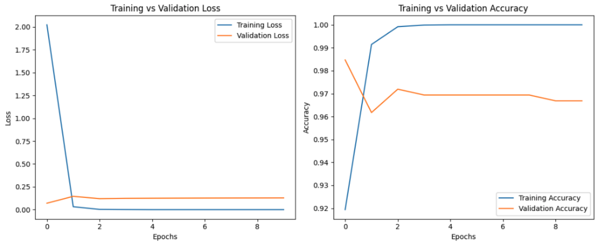

Figure 6: Pre-trained and fine-tuned VGG