IJEMD-CSAI, 2 (2) (2023), 16 – 18
https://doi.org/10.54938/ijemdcsai.2023.02.2.638

IJEMD

International Journal of Emerging Multidiciplinaries: Computer Science & Artificial Intelligence

Research Paper
Journal Homepage: www.ijemd.com
ISSN (print): 0000-0000



Stable Backpropagation in Deep Image Recognition: A VGGNet-Centered Analysis with Modern AI Perspectives

Meng Chengshuo[1] , Pan Junyu[2] , Lei Kaisong[3] , Deng Mile[4] , Ryan Chia[5], Chung Hern[6]
[1]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia
[2]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia
[3]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia
[4]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia
[5]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia
[6]School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia

ARTICLE INFO

Submitted:
15-09-2023
Revised:
30-10-2023
Accepted:
15-11-2023
Keywords:
VGGNet, backpropagation, gradient flow, convolutional neural networks, image recognition, Vision Transformers, self-supervised learning, CIFAR-10
Corresponding Author:
Meng Chengshuo

ABSTRACT

This paper explores the properties of back propagation and gradient flow of three basic CNN architectures – VGGNet, ResNet and Inception (GoogLeNet) – in image recognition. For each architecture, we explain the details of forward propagation, the calculation of categorical cross-entropy loss and the backward pass mechanisms. Training stability and transparent gradient monitoring is established in a VGGNet implementation on CIFAR-10. In addition to classical analysis, some modern AI paradigms such as Vision Transformers, self-supervised learning, neural architecture search, and foundation models are revolutionizing the field of image recognition, and are being introduced to augment the capabilities of CNN-based approaches.

1. INTRODUCTION

Another crucial field of AI is image recognition, which empowers computers to decipher images and categorize them into meaningful classes. The field has made progress from feature detection systems programmed manually to deep learning models which learn relevant features from vast amounts of data. The most popular is the Convolutional Neural Network (CNN), inspired by biological visual processing, that are "stacked" to progressively learn simple edges up to complex shapes.

CNN is trained by using forward propagation to generate predictions, a loss function to quantify error, and back propagation to update weights by computing gradients. The ability to capture the stability and efficiency of gradient flow during backprop is crucial for an effective learning ability of a model, especially for deep architectures where vanishing gradients are a major concern.

There are three representative architectures that deal with gradient flows differently: VGGNet keeps gradient flow single, straight, and simple; ResNet adds skip connections for direct gradient flow; and GoogLeNet uses multi-branch Inception modules and auxiliary classifiers. Both approaches have their own pros and cons in terms of simplicity, efficiency, and scalability to depth.

AI advances have recently given rise to Vision Transformers (ViTs) equipped with self-attention mechanisms, self-supervised learning to minimise reliance on labelled data , and automated model architecture search and foundation models such as CLIP for multimodal understanding . In this paper, we analyze the gradient flow within the three classical architectures, implement a VGGNet on CIFAR-10 and place them in the modern AI context.

2.MODEL ANALYSIS

Forward Pass

In the forward propagation, the input image will go through convolutional layers that are used to apply the learnable filters, which will generate feature maps; then, the feature maps will undergo ReLU activation for nonlinearity and pooling for spatial reduction. The operation performed at each convolutional layer is $Z(l) = W(l) * A(l-1) + b(l)$ where $A(l) = \max(0, Z(l))$.

VGGNet processes the input data in a very sequential manner, consisting of three-step 3×3 convolutions with high interpretability. ResNet introduces skip connections: $y = F(x, W) + x$, which helps to retain information throughout layers [4]. Inception uses multiple convolutions in parallel but uses different kernel sizes (1×1, 3×3, 5×5) and then concatenate outputs to create multi-scale feature extraction.

© 2023 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

Loss Function

All three architectures rely on Categorical Cross-Entropy (CCE) loss: $L = -\sum y_i \log(\hat{y}_i)$, where y_i is the one-hot true label, and $\hat{y}_i$ is the softmax-predicted probability. The loss is used as the optimization signal for backpropagation. GoogLeNet augments this with auxiliary losses at intermediate layers for enhanced gradient flow.The backward pass and gradient flow are used.Backpropagation applies the chain rule: $\partial L/\partial W(l) = (\partial L/\partial A(l)) \cdot (\partial A(l)/\partial Z(l)) \cdot (\partial Z(l)/\partial W(l))$ [2]. VGGNet has a single clear path for the propagation of its gradients, a feature known as a stacked structure. ReLU's derivative(when active) is used to prevent gradient shrinkage, and the convolutional gradient formula $\nabla KL = conv(\nabla YL, rot180(X))$ is used to ensure stable propagation.

The formula of ResNet introduces a "+1" term for the skip connections, so that the gradient of the loss function is never completely lost when the residual gradient is close to zero [4]: $\nabla xL = \nabla HL \cdot (\nabla xF(x) + 1)$. Gradients are distributed across 4 parallel branches in Inception: $\nabla xL = \sum \nabla YiL * Ki$, where auxiliary classifiers add some additional gradient injection points.

3.MODERN AI PARADIGMS

Vision Transformers

ViTs view images as sequences of patches and pass them through Transformer encoders that are based on self-attention. They model global dependencies with a large scale of pretraining[20]. Hierarchical features in Swin Transformers are introduced by the shifted-window attention for practical performance.

Self-Supervised Learning

The contrastive learning (SimCLR, BYOL) and masked image modeling (MAE) learn robust visual representations without labels. These approaches not only decrease the amount of annotation but also enrich the representation for subsequent tasks.

Reinforcement Learning for Computer Vision

NAS automates the discovery of architecture. Superior Accuracy-Efficiency trade-offs are realized by efficient scaling with compound scaling, by EfficientNet. Hardware-aware NAS guarantees deployability on resource constrained devices.

Foundation Models

Multimodal pretraining on image-text pairs has been shown to yield highly transferable representations in the context of zero-shot and few-shot visual tasks in CLIP and Flamingo.

4.THIS IS A SIMPLE IMPLEMENTATION OF VGG WITH ACCURACY ON CIFAR-10 DATASET.

A. Installation and System Design

Implementing a VGGNet on CIFAR-10 (32×32 color images, 10 classes) using TensorFlow/Keras. The network architecture is made up of three VGG blocks, each having 3x3 Conv-BatchNorm-ReLU layers followed by a max-pooling layer with filters of 64, 128 and 256 respectively. The dense flattening is replaced by Global Average Pooling and softmax followed by dropout output. Throughout the following were used: He initialization and L2 regularization.

B. Training and Results

The training was done with SGD optimizer with learning rate of 0.01. Loss and accuracy curves exhibited smooth convergence indicating stability in training and validation metrics and showing close tracking, which implied a good flow of the gradient through the single-path architecture. The final accuracy obtained was 46.88% with accuracy of 48.46% on validation set, showing that the model is not severely overfitted, but learning was well balanced. The clear gradient trajectory allowed to easily observe the norms of the gradient and diagnose training behaviour.

5.CRITICAL EVALUATION

VGGNet's single-path architecture offers the best visualisation of a gradient flow and the most explainable training dynamics, making it suitable for grasping the basics of backpropagation. It's prone to vanishing gradients in very deep configurations, however, and also has a high number of parameters. ResNet's skip connections are the most effective at maintaining gradients across deep networks, whereas Inception's multi-branch architecture is efficient in terms of computation and allows for multiple paths of gradients.

In the era of AI, each of these architectures is used as a building block. Backbones are used in ResNet, which are widely used in Vision Transformers and object detection . These fundamental concepts of the classical models also continue to guide the design of architectures for self-supervised learning, NAS and multi-modal systems .

5.CONCLUSION

This paper showed that VGGNet's one-path design ensures stable and interpretable backpropagations for image recognition, and ResNet and GoogLeNet have better gradient management for deeper networks. These basic concepts are being expanded by emerging AI paradigms such as Vision Transformers, self-supervised learning, and foundation models, establishing a complex ecosystem where CNN-based architectures are still valuable components.

6.References

Alayrac, J.-B., et al. (2022). Flamingo: A visual language model for few-shot learning. In Advances in Neural Information Processing Systems.

Alzubaidi, L., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1).

Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., & Humayun, M. (2023). A deep learning approach for atrial fibrillation classification using multi-feature time series data from ECG and PPG. Diagnostics, 13(14), Article 2442. https://doi.org/10.3390/diagnostics13142442

Chen, T., et al. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597–1607).

Dosovitskiy, A., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

Ghosh, S., Singh, A., Kavita, Jhanjhi, N. Z., Masud, M., & Aljahdali, S. (2022). SVM and KNN based CNN architectures for plant classification. Computers, Materials & Continua, 72(1), 1927–1945. https://doi.org/10.32604/cmc.2022.023414

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

He, K., et al. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000–16009).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training. In Proceedings of the 32nd International Conference on Machine Learning (pp. 448–456).

Jhanjhi, N. Z. (2025). Investigating the influence of loss functions on the performance and interpretability of machine learning models. In S. Pal & Á. Rocha (Eds.), Proceedings of the 4th International Conference on Mathematical Modeling and Computational Science (Lecture Notes in Networks and Systems, Vol. 1399). Springer. https://doi.org/10.1007/978-3-031-91005-0_43

Khalil, M. I., Humayun, M., Jhanjhi, N. Z., Talib, M. N., & Tabbakh, T. A. (2021). Multi-class segmentation of organ at risk from abdominal CT images: A deep learning approach. In S. L. Peng, S. Y. Hsieh, S. Gopalakrishnan, & B. Duraisamy (Eds.), Intelligent computing and innovation on data science (Lecture Notes in Networks and Systems, Vol. 248). Springer. https://doi.org/10.1007/978-981-16-3153-5_45

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097–1105).

Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022).

Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abdelmaboud, A. (2022). A trust-based model for secure routing against RPL attacks in Internet of Things. Sensors, 22(18), Article 7052. https://doi.org/10.3390/s22187052

Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536.

Saeed, S., Abdullah, A., Jhanjhi, N. Z., et al. (2022). New techniques for efficiently k-NN algorithm for brain tumor detection. Multimedia Tools and Applications, 81, 18595–18616. https://doi.org/10.1007/s11042-022-12271-x

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations.

Szegedy, C., et al. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–12).

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (pp. 6105–6114).

Wu, B., et al. (2019). FBNet: Hardware-aware efficient ConvNet design via differentiable NAS. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10734–10742).

Zaman, S. K. u., Jehangiri, A. I., Maqsood, T., Umar, A. I., Khan, M. A., Jhanjhi, N. Z., Shorfuzzaman, M., & Masud, M. (2022). COME-UP: Computation offloading in mobile edge computing with LSTM based user direction prediction. Applied Sciences, 12(7), Article 3312. https://doi.org/10.3390/app12073312