

Comparative Analysis of Machine Learning Models: Alexnet, VGG, Resnet, YOLO

Pham Duc An

10422002

Tran Hai Duong

10422021

Vo Thi Hong Ha

10421015

Nguyen Hoang Anh Khoa

10422037

Truong Hao Nhien

10422062

Nguyen Song Thien Phuc

10422067

Bui Duc Xuan

10422085

{@student.vgu.edu.vn}

Abstract

In this project, we conducted a comprehensive comparative analysis of prominent machine learning models, namely Alexnet, VGG, Resnet, and YOLO, with a focus on their efficacy in image recognition. Leveraging a curated dataset representative of diverse real-world scenarios with CIFAR-10, our study delved into the nuances of each model's architecture, training process, and computational requirements. Through rigorous evaluation using metrics such as accuracy, precision, and recall, our results reveal nuanced performance distinctions. Notably, Resnet demonstrated superior accuracy, VGG excelled in feature extraction, YOLO showcased real-time efficiency, and Alexnet exhibited a stable performance. These findings provide valuable insights for practitioners and researchers seeking to optimize model selection for specific applications, shedding light on the trade-offs between accuracy, computational cost, and real-time processing capabilities. Project's detailed code are provided at <https://github.com/nhientruong04/LIA-introCS-proj>.

1. Models

1.1. ResNet

ResNet is an innovative result realized by a team of Microsoft researchers during participating the ILSVRC 2015, with the ultimate grand achievement is the first place of the competition by using multiple ensembles rooted from ResNet. Its original paper [6], has gained substantial citations and is considered one of the most important contributions in the field of Computer Vision.

1.1.1 Addressed problems

Main problems With the rapid development of convolutional neural networks after the success of VGGNet [9], various networks have been created based on the idea of VGG, “the deeper, the better”. Despite multiple slight differences in terms of layers or activation functions in those models, the same technique aiming at creating a model that was “deep enough” was utilized popularly at that time. However, a new problem was raised questioning the “deepness” of any CNN models: “How deep can a model be extended?”. The problem is called the degradation phenomenon, when a deeper version of a CNN model performs worse than its shallower counterparts. This experiment was shown in the paper of Resnet, when the authors trained 2 plain models and compared their performance, as shown in 1. It is clear that the 20-layer model performed better than its deeper counterpart in both training and testing set. This graph also showed that the degradation phenomenon was not correlated with the overfitting problem [13], as the training error of the 56-layer model was higher than the 20-layer one and no divergence between the training error and the testing error was found. The authors argued that this phenomenon is caused when an already sufficiently deep model gets deeper, then its accuracy and training error would saturate earlier instead of achieving its efficient local extremum.

The pitfall of vanishing gradients The degradation phenomenon mentioned above is currently still a controversial topic, when many sources [11, 2, 4] agree that this phenomenon is rooted from the vanishing gradients problem [5, 1]. This is indeed a valid reason since the gradients would gradually decay through each individual layer in the backpropagation phase. However, the authors of ResNet argued that the vanishing gradients problem has little effect on the model causing the degradation phenomenon. “We ar-

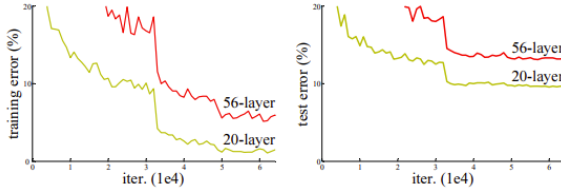


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The image was taken from the ResNet paper [6].

gue that this optimization difficulty is unlikely to be caused by vanishing gradients. These plain networks are trained with BN, which ensures forward propagated signals to have non-zero variances. We also verify that the backward propagated gradients exhibit healthy norms with BN. So neither forward nor backward signals vanish.”, stated by Kaiming He et al. in their paper [6]. Based on this argument, they argued that the degradation phenomenon was indeed a natural dilemma arising from the action of deepening any CNN models. This argument directly opposes the effect of the vanishing gradients problem on deep models, which would require that more proof and experiments must be conducted to approve any sides. However, despite the different reasoning of the degradation phenomenon, one of the same goals of any models created till now is avoiding this problem.

1.1.2 Architecture

Skip connections Mathematically, defining a residual block with a skip connection as:

$$y = \gamma(x) + x$$

With $\gamma(x)$ represents a generalized convolutional mapping with an activation function. The equation strictly illustrates combining input x with $\gamma(x)$ through addition expression. A skip connection is a pathway for any input feature map to flow to the output of the next layer. The intuition behind this technique resembles the process of a person reviewing a picture one more time that has not seen for a long time. Adding the input to the output feature map is believed to make the model “remember” its input, since any processing or calculation that input went through may make the output completely unrelated. Moreover, many sources of information about ResNet besides the authors agree that these skip connections will alleviate the vanishing gradients problem, hence could help the model to have a greater amount of layers.

The general architecture Since skip connections can be used as an identity mapping in any CNN models, the proposed ResNet would have many layers, from 34 to 151 lay-

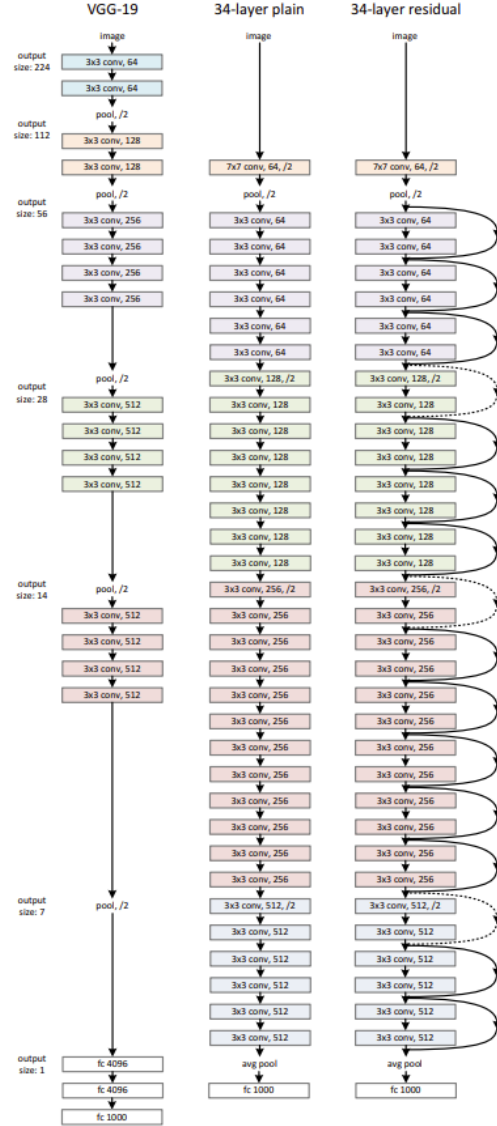


Figure 2. Architecture of ResNet compared to plain models. The image was taken from the ResNet paper [6].

ers preferably. The skip connections are utilized in each layer, as shown in 2. The proposed ResNet model resembles the VGGNet in terms of using small kernel size for convolutional layers and stacking a sufficient amount of these layers. The skip connections can be directly used when the input and output are of the same dimensions (solid line shortcuts in 2). When the dimensions increase (dotted line shortcuts in 2), there are 2 options: (A) The shortcut still performs identity mapping, with extra zero entries padded for increasing dimensions. This option introduces no extra parameter; (B) The projection shortcut is used to match dimensions (done by 1×1 convolutions). For both options, when the shortcuts go across feature maps of two sizes, they

are performed with a stride of 2.

1.1.3 Applications

The invention of ResNet sparks the utilization of that same structure, especially the skip connections or residual blocks, in many CNN models not only in the Computer Vision field but also a greater, older field which is Natural Language Processing. Some impactful architectures developed based on ResNet are ResNeXt [12] and DenseNet [7], where residual blocks and skip connections are implemented differently to tackle the weaknesses of ResNet. In addition, those skip connections (or preferably called identity mapping), are utilized in Transformer [10], which is an evolutionary architecture that outperformed any forms of CNN in NLP. The same architecture was also widely used to create different types of Transformer models in the Computer Vision field, such as ViT [3] or Swin [8]. The skip connections used in Transformer sequentially made the Attention blocks in this model a new type of residual blocks.

References

- [1] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
- [2] Datagen. Resnet: The basics and 3 resnet extensions, May 2023.
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [4] GeeksforGeeks. Residual networks (resnet) - deep learning, Jan 2023.
- [5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- [7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks, 2018.
- [8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.
- [9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.
- [11] visoai. Deep residual networks (resnet, resnet50) - 2024 guide, Nov 2023.
- [12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks, 2017.
- [13] X. Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, feb 2019.