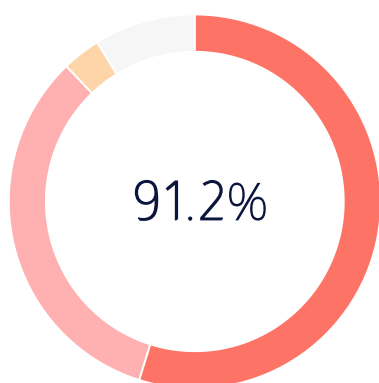


Analysis Report

Plagiarism Detection and AI Detection Report

He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Plagiarism Detection

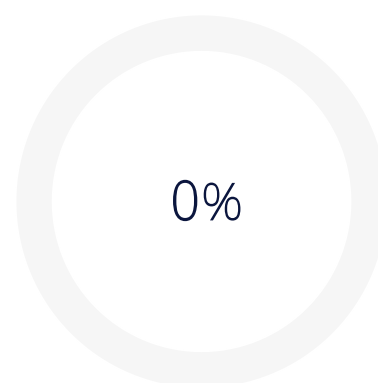


Plagiarism types	Text coverage	Words
Identical	54.8%	4,182
Minor Changes	33%	2,519
Paraphrased	3.3%	254

Excluded

Omitted Words	0
---------------	---

AI Detection

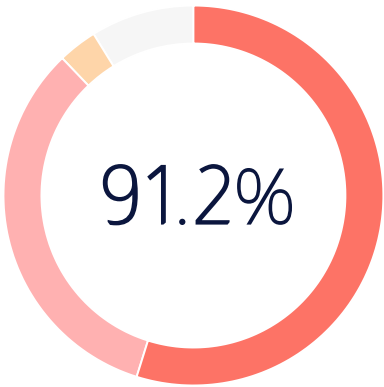


	Text coverage	Words
AI Text	0%	0
Human Text	100%	7,627

Excluded

Omitted Words	0
---------------	---

Plagiarism



Results (75)

*Results may not appear because the feature has been disabled.

 Private Cloud Hub 0	 Shared Data Hub 0	 Filtered / Excluded 0
 Internet Sources 71	 AI Source Match 4	 Current Batch 0

Plagiarism types	Text coverage	Words
 Identical	54.8%	4,182
 Minor Changes	33%	2,519
 Paraphrased	3.3%	254
Excluded		
 Omitted Words		0

About Plagiarism Detection

Our AI-powered plagiarism scans offer three layers of text similarity detection: Identical, Minor Changes, and Paraphrased. Based on your scan settings we also provide insight on how much of the text you are not scanning for plagiarism (Omitted words).

Identical

One to one exact word matches. [Learn more](#)

Minor Changes

Words that hold nearly the same meaning but have a change to their form (e.g. “large” becomes “largely”). [Learn more](#)

Paraphrased

Different words that hold the same meaning that replace the original content (e.g. 'large' becomes 'big') [Learn more](#)

Omitted Words

The portion of text that is not being scanned for plagiarism based on the scan settings. (e.g. the 'Ignore quotations' setting is enabled and the document is 20% quotations making the omitted words percentage 20%) [Learn more](#)

Copyleaks Shared Data Hub

Our Shared Data Hub is a collection of millions of user-submitted documents that you can utilize as a scan resource and choose whether or not you would like to submit the file you are scanning into the Shared Data Hub. [Learn more](#)

Filtered and Excluded Results

The report will generate a complete list of results. There is always the option to exclude specific results that are not relevant. Note, by unchecking certain results, the similarity percentage may change. [Learn more](#)

Current Batch Results

These are the results displayed from the collection, or batch, of files uploaded for a scan at the same time. [Learn more](#)

🔍 Plagiarism Detection Results: (75)

🌐 Sources – Compute-Structure AI History 87.6%
https://ai-papers.xuzou.cn/sources/2015-resnet_2015.txt
Source notes and reports Sources ← Home Deep Residual Lea...

🌐 Deep Residual Learning for Image Recognition 46.1%
https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/he_deep_residual_learning_cvpr_2016...
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Microsoft Research Jian Sun {kahe, v-xiangz, v-shre...

🌐 1512.03385 43.8%
<https://arxiv.org/pdf/1512.03385>
Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Microsoft Research Jian Sun Deeper neural networks ...

🌐 1512.03385.pdf 43.8%
<https://arxiv.org/pdf/1512.03385>
Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Microsoft Research Jian Sun Deeper neural networks ...

🌐 1-resnet_paper.pdf 43.8%
<https://storage.yandexcloud.net/lms-vault/private/2/courses/2016-summer/spb-deep-learning/materials/1-re...>
Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Microsoft Research Jian Sun Deeper neural networks ...

🌐 day10_annotated.pdf 43.6%
https://www.khoury.northeastern.edu/home/hand/teaching/cs7150-spring-2021/day10_annotated.pdf
1) Explain all of the terms in the optimization algorithm presented in Section 5 2) The AlexNet paper used a learning rate schedule wher...

🌐 Deep Residual Learning for Image Recognition 39.1%
<https://arxiv.org/html/1512.03385v1>
arXiv is now an independent nonprofit! Learn more License: arXiv.org perpetual non-exclusive licens...

🌐 [1512.03385] Deep Residual Learning for Image Recognition 36.5%
<https://arxiv.labs.arxiv.org/html/1512.03385>
Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun Affiliation: Microsoft Res...

 Deep Residual Learning for Image Recognition - ResNet Overview - Studocu https://www.studocu.com/in/document/mahatma-gandhi-institute-of-technology/deeplearning/deep-residual... Deep Residual Learning for Image Recognition Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun Microsoft Research {kahe, v-xiangz, v...	23.2%
 (PDF) Image Recognition using Deep Residual Learning https://www.researchgate.net/publication/361362336_image_recognition_using_deep_residual_learning Home Agricultural Science Residue ArticlePDF AvailableImage Recognition using Deep Residual Learning May 2022 Authors: G Ravi K...	14.4%
 易 AI - ResNet 论文深度讲解 MakeOptim https://makeoptim.com/deep-learning/yiai-paper-resnet/ ...	12.7%
 Deep Residual Learning for Image Recognition(ResNet)(翻譯) - HackMD https://hackmd.io/@shaoechen/sy_e1mceu # Deep Residual Learning for Image Recognition(ResNet)(翻譯) ##### tags: `CNN` `論文翻譯` `deeplearning` >[name=Shaoe.chen] [tim...	9.4%
 dlc-slides-6-5-residual-networks.pdf https://fleuret.org/dlc/materials/dlc-slides-6-5-residual-networks.pdf Deep learning 6.5. Residual networks François Fleuret https://fleuret.org/dlc/ The "Highway networks" by Srivastava et al. (2015) use t...	7.1%
 ResNet 논문 https://bablove127.tistory.com/m/entry/resnet-%eb%85%bc%eb%ac%b8 ...	5.9%
 2016 DeepResidualLearningforImageRec - GM-RKB https://www.gabormelli.com/rkb/2016_deepresiduallearningforimagerec 2016 DeepResidualLearningforImageRec From GM-RKB Jump to navigation Jump to search (He et al., 2016...	5.7%
 Medium https://liyini2015.medium.com/cnn-model-residual-net-e82b6a0a379d SitemapOpen in app Sign in Medium Logo Write Search Sign in Li Yin Creator of AdaFlow: https://github.com/SylphAI...	5.3%

 Medium https://medium.com/@rchhetry5/paper-review-on-deep-residual-learning-from-image-recognition-a4e9488b1...	4.2%
SitemapOpen in app Sign in Medium Logo Write Search Sign in Towards Data Science Resnet Data Science Rev...	
 The Star Also Rises: ResNet (三) : Illustrated http://hemingwang.blogspot.com/2021/03/resnetillustrated.html	3.3%
skip to main skip to sidebar The Star Also Rises Sunday, May 16, 2021 ...	
 The Star Also Rises: May 2021 http://hemingwang.blogspot.com/2021/05/	3.3%
skip to main skip to sidebar The Star Also Rises Sunday, May 30, 2021 ...	
 Going Deep http://comp6248.ecs.soton.ac.uk/handouts/deepnetworks-handouts.pdf	3%
Kate Farrahi	
Approximate Functions Going Deep Kate Farrahi Vision, Learning and Control University of Southampton Kate Farrahi Deep Learning 2 / 2...	
 Understanding ResNets dhruv's space https://dhruvs.space/posts/understanding-resnets/	2.3%
I'm currently enrolled in fastai's Deep Learning MOOC (version 3), and loving it so far. It's only been 2 lectures as of today, but folks...	
 ResNet (34, 50, 101): Residual CNNs for Image Classification Tasks https://neurohive.io/en/popular-networks/resnet/	2.3%
Skip to content ResNet i...	
 Medium https://medium.com/@aschandinip/resnet-34-50-101-what-actually-it-is-c63da24ba695	2.3%
SitemapOpen in app Sign in Medium Logo Write Search Sign in ResNet (34, 50, 101)...what actually it is ? ch...	
 CVPR 2016 Open Access Repository https://openaccess.thecvf.com/content_cvpr_2016/html/he_deep_residual_learning_cvpr_2016_paper.html	2.2%
CVPR 2016 open access These CVPR 2016 papers are the Open Access versions, provided by the Computer Vision Foundation. Except ...	

 Resnet ML_101 https://ztlevi.gitbook.io/ml-101/neural-networks/resnet For the complete documentation index, see lms.txt. This page is also available as Markdown. Increasing network depth does not work ...	2.1%
 Deep Residual Learning for Image Recognition - Microsoft Research https://www.microsoft.com/en-us/research/publication/deep-residual-learning-for-image-recognition/ Skip to main content Research ...	2.1%
 Feature Evaluation of Deep Convolutional Neural Networks for Object Recogniti... https://arxiv.org/html/1509.07627v1 arXiv is now an independent nonprofit! Learn more License: arXiv.org perpetual non-exclusive licens...	1.9%
 [1512.03385] Deep Residual Learning for Image Recognition https://arxiv.org/abs/1512.03385 Skip to main content Press Enter to search · Advanced search Com...	1.8%
 [English] [Paper Review] Deep Residual Learning for Image Recognition (ResNet) https://eugene129.tistory.com/12 본문 바로가기 ...	1.6%
 Factorized Bilinear Models for Image Recognition https://arxiv.org/html/1611.05709v2 arXiv is now an independent nonprofit! Learn more License: arXiv.org perpetual non-exclusive licens...	1.5%
 Medium https://medium.com/analytics-vidhya/deep-residual-learning-for-image-recognition-resnet-paper-explained-2... SitemapOpen in app Sign in Medium Logo Write Search Sign in Analytics Vidhya · Analytics Vidhya is a commu...	1.5%
 CSE5194: ResNet and ResNeXt https://drago1234.github.io/about_me/pdf/cs5194_resnet_v2.0.pdf Dong, Zhengqi CSE5194: ResNet and ResNeXt The Ohio State University Zhengqi Dong E-mail: dong.760@osu.edu Date: 09/20/2019 PartI: (ResNet) Deep Residu...	1.3%

 Aldventure - ResNet - Deep Residual Learning for Image Recognition http://aidventure.es/posts/resnet/ ResNet's main contribution is the introduction of the residual block, which enables the creation of much deeper neural networks wi...	1.2%
 Deep Residual Learning for Image Recognition https://bibbase.org/network/publication/he-zhang-ren-sun-deepresiduallearningforimagerecognition Deep Residual Learning for Image Recognition. He, K., Zhang, X., Ren, S., & Sun, J. In 2016 IEEE Conference on Computer Vision and Pa...	1.2%
 [Paper Review] Deep Residual Learning for Image Recognition (ResNet) https://eugene129.tistory.com/10 본문 바로가기 ...	1.2%
 Medium https://medium.com/@manthasaigopal/resnet-paper-explained-090f71ed6e51 SitemapOpen in app Sign in Medium Logo Write Search Sign in Deep Learning Artificial Intelligence Neural Ne...	1%
 International Journal of Scientific Research in Computer Science, Engineering... https://ijsrcseit.com/paper/cseit11953130.pdf TechnoScience Academy;http://technoscienceacademy.com International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT Volume 5 Is...	1%
 Tag Prediction at Flickr: A View from the Darkroom https://arxiv.org/html/1612.01922v3 arXiv is now an independent nonprofit! Learn more License: arXiv.org perpetual non-exclusive licens...	1%
 Medium https://medium.com/@nurlan.imanov/resnet-deep-residual-learning-for-image-recognition-paper-explanation-... SitemapOpen in app Sign in Medium Logo Write Search Sign in Nurlan Imanov Data Engineer www.linkedin.com/in/nurlan...	0.9%
 (PDF) A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond https://www.researchgate.net/publication/369760111_a_comprehensive_review_of_yolo_from_yolov1_to_yolov8_... AI Source Match Home Mental Processes Psychology Comprehension PreprintPDF AvailableA Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyon...	0.9%

 DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection 0.9%
<https://www.ee.cuhk.edu.hk/~xgwang/papers/deepidnetcvpr15.pdf>

Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Cl
DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Lu...

 DeepIDNet_CVPR15.pdf 0.9%
https://wlouyang.github.io/papers/deepidnet_cvpr15.pdf

DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Lu...

 (PDF) Development of Skip Connection in Deep Neural Networks for Computer Vis... 0.8%
https://www.researchgate.net/publication/383609186_development_of_skip_connection_in_deep_neural_netwo...

AI Source Match
PreprintPDF AvailableDevelopment of Skip Connection in Deep Neural Networks for Computer Vision and Medical Image Analysis: A Survey...

 Embodied AI 101 | Transcript: Deep Residual Learning for Image Recognition (R... 0.8%
<https://embodied.transistor.fm/episodes/deep-residual-learning-for-image-recognition-resnet/transcript>

← Previous · All Episodes · Next → Episode ...

 content 0.7%
<https://cf003.cdn.4science.cloud/4s-assetstore-prod--cornell--ecommons/91/90/86/9190861331511065908182...>

TRAINING PARADIGMS FOR DEEP RESIDUAL NETWORKS A Thesis Presented to the Faculty of the Graduate School of Cornell University in Partial ...

 Biological direct-shortcut deep residual learning for sparse visual processin... 0.7%
<https://www.nature.com/articles/s41598-024-62756-y>

Skip to main content Thank you for visiting nature.com. You are using a browser version with limited supp...

 GitHub - Bao-Jiarong/ResNet: This repository includes ResNet18, ResNet34, Res... 0.6%
<https://github.com/bao-jiarong/resnet>

Skip to content ...

 Medium 0.6%
https://medium.com/@pranav.bhagwat_36/understanding-the-residual-learning-framework-for-computer-visio...

SitemapOpen in app Sign in Medium Logo Write Search Sign in Computer Vision Understanding the Residual Lea...

 Deep Residual Learning for Image Recognition: A Microsoft Research Presentati... 0.6%

<https://www.docsity.com/en/docs/deep-residual-learning-for-image-recognition-4/9594799/>

Prepare for your exams Study with the several resources on Docsity Earn points to download Earn points by helping other student...

 What is Deep Residual Learning? - Zilliz Learn 0.6%

<https://zilliz.com/learn/deep-residual-learning-for-image-recognition>

Learn AI & Machine Learning Deep Residual Learning for Image Recognition Sep 24, 202410 min read Deep residual learning solves...

 GitHub - a-martyn/resnet: An implementation of the original 'ResNet' paper in... 0.6%

<https://github.com/a-martyn/resnet>

Skip to content ...

 Notebook 0.6%

<https://csggrad.science.uoit.ca/courses/ist/notebooks/convnet.html>

Convolutional Networks¶ Faisal Qureshi <http://www.vclab.ca> Lesson Plan¶ Convolutional Networks Convolu...

 Deep policy gradient methods without batch updates, target networks, or repla... 0.5%

<https://dl.acm.org/doi/10.5555/3737916.3737942>

skip to main content ...

 Medium 0.4%

<https://medium.com/analytics-vidhya/understanding-resnet-architecture-869915cc2a98>

SitemapOpen in app Sign in Medium Logo Write Search Sign in Analytics Vidhya · Analytics Vidhya is a commu...

 Li_Dynamic_Hierarchical_Mimicking_CVPR_2020_supplemental.pdf 0.4%

https://openaccess.thecvf.com/content_cvpr_2020/supplemental/li_dynamic_hierarchical_mimicking_cvpr_2020...

Supplementary Materials for "Dynamic Hierarchical Mimicking Towards Consistent Optimization Objectives" Duo Li Qifeng Chen The Hong Kong ...

 Medium 0.4%

<https://medium.com/@romybuch22/in-between-resnets-and-densenets-b12e4ad0025a>

SitemapOpen in app Sign in Medium Logo Write Search Sign in Artificial Intelligence Deep Learning Machine L...

 (PDF) Deep Learning in Computer Vision: A Critical Review https://www.researchgate.net/publication/385259651_deep_learning_in_computer_vision_a_critical_review ArticlePDF AvailableDeep Learning in Computer Vision: A Critical Review April 2024 Journal of Electrical Systems 20(3):5540-5562 D...	0.4%
 Understanding Deep Residual Networks https://shuzhanfan.github.io/2018/11/resnet/ sf. Blo...	0.4%
 MACVI2023.pdf https://elib.dlr.de/202238/1/macvi2023.pdf AI Source Match arXiv:2311.14762v1 [cs.CV] 23 Nov 2023 2nd Workshop on Maritime Computer Vision (MaCVi) 2024: Challenge Results Benjamin Kiefer1, Lojze...	0.3%
 Effect of Convolutional Depth on Image Recognition Performance: VGG vs. ResNe... https://arxiv.org/html/2602.13298v1 arXiv is now an independent nonprofit! Learn more License: arXiv.org perpetual non-exclusive licens...	0.3%
 Learning Spatio-Temporal Features With 3D Residual Networks for Action Recogn... https://openaccess.thecvf.com/content_iccv_2017_workshops/papers/w44/hara_learning_spatio-temporal_feat... Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National ...	0.3%
 Training error (left) and test error (right) on CIFAR-10 with 20-layer... D... https://www.researchgate.net/figure/training-error-left-and-test-error-right-on-cifar-10-with-20-layer-and-56... Figure - available from: Scientific ReportsThis content is subject to copyright. Terms and conditions apply. Download View publi...	0.3%
 Medium https://medium.com/@ngiengkianyew/residual-connection-intuition-c8cf011170d SitemapOpen in app Sign in Medium Logo Write Search Sign in Ngieng Kianyew Just another Data Scientist ...	0.3%
 Highway Networks, the first working really deep feedforward neural networks w... https://people.idsia.ch/~juergen/highway-networks.html . Jürgen Schmidhuber (2015, updated 2025) Pronounce: You_again Shmidhoobuh AI Blog @8#8203;SchmidhuberAI Highway Networks...	0.2%

 (PDF) Human pose estimation and its application to action recognition: A survey https://www.researchgate.net/publication/349853529_human_pose_estimation_and_its_application_to_action_r... ArticlePDF AvailableHuman pose estimation and its application to action recognition: A survey April 2021 Journal of Visual Communica...	0.2%
 Kolmogorov-Arnold Convolutions: Design Principles and Empirical Studies https://arxiv.org/html/2407.01092v1 AI Source Match arXiv is now an independent nonprofit! Learn more License: CC BY 4.0 arXiv:2407.01092v1 [cs.C...	0.2%
 Network parameters of the DB-ResNet. Building blocks are shown in... Downlo... https://www.researchgate.net/figure/network-parameters-of-the-db-resnet-building-blocks-are-shown-in-bra... Figure - uploaded by Renchao JinContent may be subject to copyright. Download View publication Copy reference Copy caption Embed ...	0.2%
 Medium https://smlra-kjsce.medium.com/whats-the-res-in-resnet-part-ii-96612490cb8e?source=author_recirc-----8a3a... SitemapOpen in app Sign in Medium Logo Write Search Sign in Machine Learning Resnet Deep Learning Artifi...	0.2%
 ResNet: Enabling Deep Convolutional Neural Networks through Residual Learning https://arxiv.org/html/2510.24036v1 arXiv is now an independent nonprofit! Learn more License: CC BY 4.0 arXiv:2510.24036v1 [cs.C...	0.2%
 (PDF) Training Matters: Unlocking Potentials of Deeper Graph Convolutional Ne... https://www.researchgate.net/publication/375756826_training_matters_unlocking_potentials_of_deeper_graph... Home Organizational Psychology Training ArticlePDF AvailableTraining Matters: Unlocking Potentials of Deeper Graph Convolutional Ne...	0.2%
 GitHub - dnlcrl/deep-residual-networks-pyfunt: Python implementation of 'Deep... https://github.com/dnlcrl/deep-residual-networks-pyfunt Skip to content ...	0.2%
 (PDF) Comparative analysis of VGG, ResNet, and GoogLeNet architectures evalua... https://www.researchgate.net/publication/378739119_comparative_analysis_of_vgg_resnet_and_googlenet_arch... ArticlePDF AvailableComparative analysis of VGG, ResNet, and GoogLeNet architectures evaluating performance, computational efficienc...	0.2%

 Adaptive skip connections improve training - Nick Ryan 0.2%
<https://nickcdryan.com/2024/05/24/adaptive-skip-connections-improve-training/>
Skip to content Adaptive skip connections improve training SUMMARY Applying a single, linear, learnable weight to the skip ...

 Medium 0.2%
<https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d>
SitemapOpen in app Sign in Medium Logo Write Search Sign in Siddhesh Bangar ✨ Build Intelligent systems to solve c...

 ShuffleFaceNet: A Lightweight Face Architecture for Efficient and Highly-Accu... 0.2%
https://openaccess.thecvf.com/content_iccvw_2019/papers/lsr/martinez-diaz_shufflefacenet_a_lightweight_fa...
Yoanna Martindez-Diaz, Luis S. Luevano, Heydi Mendez-Vazquez, Miguel Nicolas-Diaz, Leonardo Chang, Miguel Gonz.
ShuffleFaceNet: A Lightweight Face Architecture for Efficient and Highly-Accurate Face Recognition Yoanna Martínez-Díaz, Heydi Méndez-V...

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

hkahe, v-xiangz, v-shren, jiansun@microsoft.com

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [40] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers.

The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions¹, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

1. Introduction

Deep convolutional neural networks [22, 21] have led to a series of breakthroughs for image classification [21, 49, 39]. Deep networks naturally integrate low/mid/high-level features [49] and classifiers in an end-to-end multi-layer fashion, and the “levels” of features can be enriched by the number of stacked layers (depth). Recent evidence [40, 43] reveals that network depth is of crucial importance, and the leading results [40, 43, 12, 16] on the challenging ImageNet dataset [35] all exploit “very deep” [40] models, with a depth of sixteen [40] to thirty [16]. Many other non-trivial visual recognition tasks [7, 11, 6, 32, 27] have also

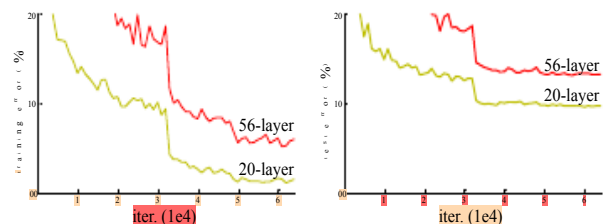


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

greatly benefited from very deep models.

Driven by the significance of depth, a question arises: Is learning better networks as easy as stacking more layers? An obstacle to answering this question was the notorious problem of vanishing/exploding gradients [14, 1, 8], which hamper convergence from the beginning. This problem, however, has been largely addressed by normalized initialization [23, 8, 36, 12] and intermediate normalization layers [16], which enable networks with tens of layers to start converging for stochastic gradient descent (SGD) with back-propagation [22].

When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error, as reported in [10, 41] and thoroughly verified by our experiments. Fig. 1 shows a typical example.

The degradation (of training accuracy) indicates that not all systems are similarly easy to optimize. Let us consider a shallower architecture and its deeper counterpart that adds more layers onto it. There exists a solution by construction to the deeper model: the added layers are identity mapping, and the other layers are copied from the learned shallower model. The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart. But experiments show that our current solvers on hand are unable to find solutions that

¹<http://image-net.org/challenges/LSVRC/2015/> and <http://mscoco.org/dataset/#detections-challenge2015>.

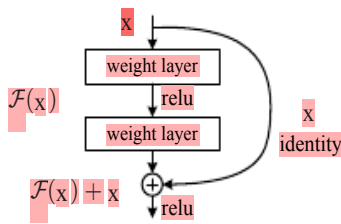


Figure 2. Residual learning: a building block.

are comparably good or better than the constructed solution (or unable to do so in feasible time).

In this paper, we address the degradation problem by introducing a deep residual learning framework. Instead of hoping each few stacked layers directly fit a desired underlying mapping, we explicitly let these layers fit a residual mapping. Formally, denoting the desired underlying mapping as $\mathcal{H}(x)$, we let the stacked nonlinear layers fit another mapping of $\mathcal{F}(x) := \mathcal{H}(x) - x$. The original mapping is recast into $\mathcal{F}(x) + x$. We hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

The formulation of $\mathcal{F}(x) + x$ can be realized by feedforward neural networks with “shortcut connections” (Fig. 2). Shortcut connections [2, 33, 48] are those skipping one or more layers. In our case, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers (Fig. 2). Identity shortcut connections add neither extra parameter nor computational complexity. The entire network can still be trained end-to-end by SGD with backpropagation, and can be easily implemented using common libraries (e.g., Caffe [19]) without modifying the solvers.

We present comprehensive experiments on ImageNet [35] to show the degradation problem and evaluate our method. We show that: 1) Our extremely deep residual nets are easy to optimize, but the counterpart “plain” nets (that simply stack layers) exhibit higher training error when the depth increases; 2) Our deep residual nets can easily enjoy accuracy gains from greatly increased depth, producing results substantially better than previous networks.

Similar phenomena are also shown on the CIFAR-10 set [20], suggesting that the optimization difficulties and the effects of our method are not just akin to a particular dataset. We present successfully trained models on this dataset with over 100 layers, and explore models with over 1000 layers.

On the ImageNet classification dataset [35], we obtain excellent results by extremely deep residual nets. Our 152-layer residual net is the deepest network ever presented on ImageNet, while still having lower complexity than VGG nets [40]. Our ensemble has 3.57% top-5 error on the

ImageNet test set, and won the 1st place in the ILSVRC 2015 classification competition. The extremely deep representations also have excellent generalization performance on other recognition tasks, and lead us to further win the 1st places on: ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in ILSVRC & COCO 2015 competitions. This strong evidence shows that the residual learning principle is generic, and we expect that it is applicable in other vision and non-vision problems.

2. Related Work

Residual Representations. In image recognition, VLAD [18] is a representation that encodes by the residual vectors with respect to a dictionary, and Fisher Vector [30] can be formulated as a probabilistic version [18] of VLAD. Both of them are powerful shallow representations for image retrieval and classification [4, 47]. For vector quantization, encoding residual vectors [17] is shown to be more effective than encoding original vectors.

In low-level vision and computer graphics, for solving Partial Differential Equations (PDEs), the widely used Multigrid method [3] reformulates the system as subproblems at multiple scales, where each subproblem is responsible for the residual solution between a coarser and a finer scale. An alternative to Multigrid is hierarchical basis preconditioning [44, 45], which relies on variables that represent residual vectors between two scales. It has been shown [3, 44, 45] that these solvers converge much faster than standard solvers that are unaware of the residual nature of the solutions. These methods suggest that a good reformulation or preconditioning can simplify the optimization.

Shortcut Connections. Practices and theories that lead to shortcut connections [2, 33, 48] have been studied for a long time. An early practice of training multi-layer perceptrons (MLPs) is to add a linear layer connected from the network input to the output [33, 48]. In [43, 24], a few intermediate layers are directly connected to auxiliary classifiers for addressing vanishing/exploding gradients. The papers of [38, 37, 31, 46] propose methods for centering layer responses, gradients, and propagated errors, implemented by shortcut connections. In [43], an “inception” layer is composed of a shortcut branch and a few deeper branches.

Concurrent with our work, “highway networks” [41, 42] present shortcut connections with gating functions [15]. These gates are data-dependent and have parameters, in contrast to our identity shortcuts that are parameter-free. When a gated shortcut is “closed” (approaching zero), the layers in highway networks represent non-residual functions. On the contrary, our formulation always learns residual functions; our identity shortcuts are never closed, and all information is always passed through, with additional residual functions to be learned. In addition, high-

way networks have not demonstrated accuracy gains with extremely increased depth (e.g., over 100 layers).

3. Deep Residual Learning

3.1. Residual Learning

Let us consider $\mathcal{H}(\mathbf{x})$ as an underlying mapping to be fit by a few stacked layers (not necessarily the entire net), with $\mathbf{x}$ denoting the inputs to the first of these layers. If one hypothesizes that multiple nonlinear layers can asymptotically approximate complicated functions², then it is equivalent to hypothesize that they can asymptotically approximate the residual functions, i.e., $\mathcal{H}(\mathbf{x}) - \mathbf{x}$ (assuming that the input and output are of the same dimensions). So rather than expect stacked layers to approximate $\mathcal{H}(\mathbf{x})$, we explicitly let these layers approximate a residual function $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$. The original function thus becomes $\mathcal{F}(\mathbf{x}) + \mathbf{x}$. Although both forms should be able to asymptotically approximate the desired functions (as hypothesized), the ease of learning might be different.

This reformulation is motivated by the counterintuitive phenomena about the degradation problem (Fig. 1, left). As we discussed in the introduction, if the added layers can be constructed as identity mappings, a deeper model should have training error no greater than its shallower counterpart. The degradation problem suggests that the solvers might have difficulties in approximating identity mappings by multiple nonlinear layers. With the residual learning reformulation, if identity mappings are optimal, the solvers may simply drive the weights of the multiple nonlinear layers toward zero to approach identity mappings.

In real cases, it is unlikely that identity mappings are optimal, but our reformulation may help to precondition the problem. If the optimal function is closer to an identity mapping than to a zero mapping, it should be easier for the solver to find the perturbations with reference to an identity mapping, than to learn the function as a new one. We show by experiments (Fig. 7) that the learned residual functions in general have small responses, suggesting that identity mappings provide reasonable preconditioning.

3.2. Identity Mapping by Shortcuts

We adopt residual learning to every few stacked layers. A building block is shown in Fig. 2. Formally, in this paper we consider a building block defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}; \mathbb{W}_i \langle \rangle) + \mathbf{x} \quad (1)$$

Here $\mathbf{x}$ and $\mathbf{y}$ are the input and output vectors of the layers considered. The function $\mathcal{F}(\mathbf{x}; \mathbb{W}_i \langle \rangle)$ represents the residual mapping to be learned. For the example in Fig. 2 that has two layers, $\mathcal{F} = W_2 \sigma(W_1 \mathbf{x})$ in which σ denotes

ReLU [29] and the biases are omitted for simplifying notations. The operation $\mathcal{F} + \mathbf{x}$ is performed by a shortcut connection and element-wise addition. We adopt the second nonlinearity after the addition (i.e., $\sigma(\mathbf{y})$, see Fig. 2).

The shortcut connections in Eqn.(1) introduce neither extra parameter nor computation complexity. This is not only attractive in practice but also important in our comparisons between plain and residual networks. We can fairly compare plain/residual networks that simultaneously have the same number of parameters, depth, width, and computational cost (except for the negligible element-wise addition).

The dimensions of $\mathbf{x}$ and $\mathcal{F}$ must be equal in Eqn.(1). If this is not the case (e.g., when changing the input/output channels), we can perform a linear projection W_s by the shortcut connections to match the dimensions:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}; \mathbb{W}_i \langle \rangle) + W_s \mathbf{x} \quad (2)$$

We can also use a square matrix W_s in Eqn.(1). But we will show by experiments that the identity mapping is sufficient for addressing the degradation problem and is economical, and thus W_s is only used when matching dimensions.

The form of the residual function $\mathcal{F}$ is flexible. Experiments in this paper involve a function $\mathcal{F}$ that has two or three layers (Fig. 5), while more layers are possible. But if $\mathcal{F}$ has only a single layer, Eqn.(1) is similar to a linear layer: $\mathbf{y} = W_1 \mathbf{x} + \mathbf{x}$, for which we have not observed advantages.

We also note that although the above notations are about fully-connected layers for simplicity, they are applicable to convolutional layers. The function $\mathcal{F}(\mathbf{x}; \mathbb{W}_i \langle \rangle)$ can represent multiple convolutional layers. The element-wise addition is performed on two feature maps, channel by channel.

3.3. Network Architectures

We have tested various plain/residual nets, and have observed consistent phenomena. To provide instances for discussion, we describe two models for ImageNet as follows.

Plain Network. Our plain baselines (Fig. 3, middle) are mainly inspired by the philosophy of VGG nets [40] (Fig. 3, left). The convolutional layers mostly have 3×3 filters and follow two simple design rules: (i) for the same output feature map size, the layers have the same number of filters; and (ii) if the feature map size is halved, the number of filters is doubled so as to preserve the time complexity per layer. We perform downsampling directly by convolutional layers that have a stride of 2. The network ends with a global average pooling layer and a 1000-way fully-connected layer with softmax. The total number of weighted layers is 34 in Fig. 3 (middle).

It is worth noticing that our model has fewer filters and lower complexity than VGG nets [40] (Fig. 3, left). Our 34-layer baseline has 3.6 billion FLOPs (multiply-adds), which is only 18% of VGG-19 (19.6 billion FLOPs).

²This hypothesis, however, is still an open question. See [28].

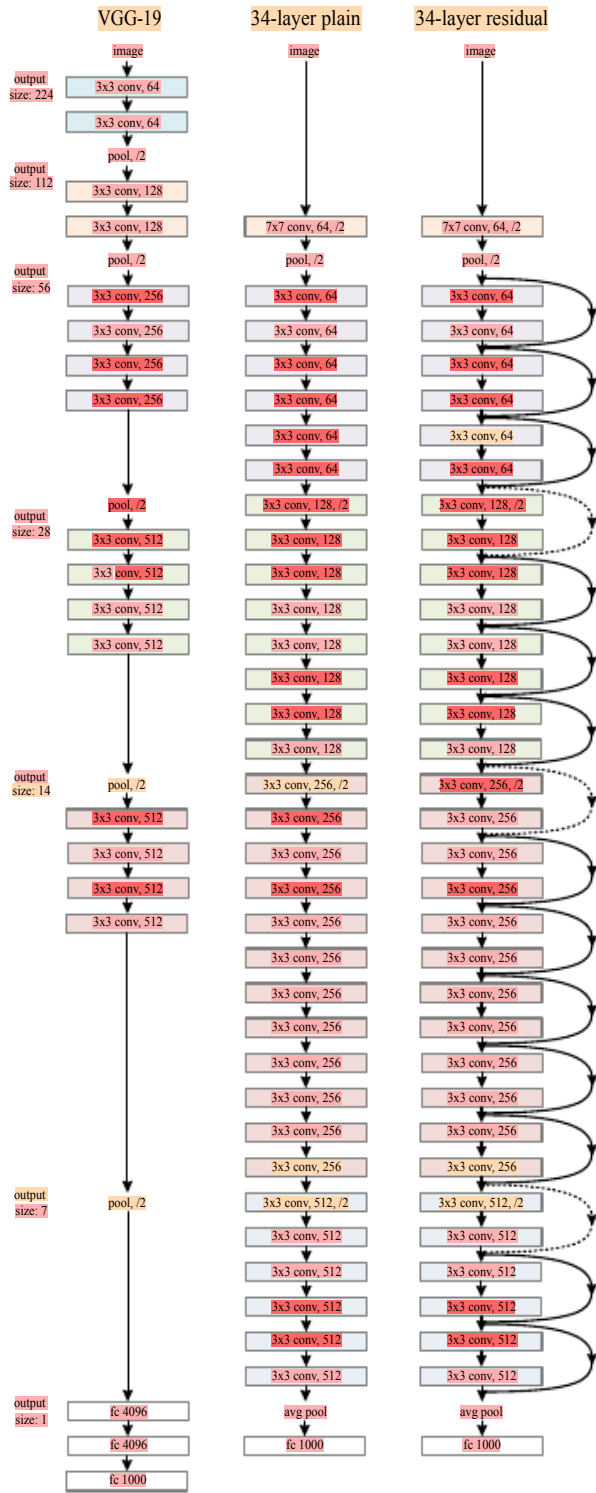


Figure 3. Example network architectures for ImageNet. Left: the VGG-19 model [40] (19.6 billion FLOPs) as a reference. Middle: a plain network with 34 parameter layers (3.6 billion FLOPs). Right: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions. Table 1 shows more details and other variants.

Residual Network. Based on the above plain network, we insert shortcut connections (Fig. 3, right) which turn the network into its counterpart residual version. The identity shortcuts (Eqn.(1)) can be directly used when the input and output are of the same dimensions (solid line shortcuts in Fig. 3). When the dimensions increase (dotted line shortcuts in Fig. 3), we consider two 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 in Eqn.(2) 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.

3.4. Implementation

Our implementation for ImageNet follows the practice in [21, 40]. The image is resized with its shorter side randomly sampled in $[256, 480]$ for scale augmentation [40]. A 224×224 crop is randomly sampled from an image or its horizontal flip, with the per-pixel mean subtracted [21]. The standard color augmentation in [21] is used. We adopt batch normalization (BN) [16] right after each convolution and before activation, following [16]. We initialize the weights as in [12] and train all plain/residual nets from scratch. We use SGD with a mini-batch size of 256. The learning rate starts from 0.1 and is divided by 10 when the error plateaus, and the models are trained for up to 60×10^4 iterations. We use a weight decay of 0.0001 and a momentum of 0.9. We do not use dropout [13], following the practice in [16].

In testing, for comparison studies we adopt the standard 10-crop testing [21]. For best results, we adopt the fully-convolutional form as in [40, 12], and average the scores at multiple scales (images are resized such that the shorter side is in $[224, 256, 384, 480, 640]$).

4. Experiments

4.1. ImageNet Classification

We evaluate our method on the ImageNet 2012 classification dataset [35] that consists of 1000 classes. The models are trained on the 1.28 million training images, and evaluated on the 50k validation images. We also obtain a final result on the 100k test images, reported by the test server. We evaluate both top-1 and top-5 error rates.

Plain Networks. We first evaluate 18-layer and 34-layer plain nets. The 34-layer plain net is in Fig. 3 (middle). The 18-layer plain net is of a similar form. See Table 1 for detailed architectures.

The results in Table 2 show that the deeper 34-layer plain net has higher validation error than the shallower 18-layer plain net. To reveal the reasons, in Fig. 4 (left) we compare their training/validation errors during the training procedure. We have observed the degradation problem - the

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112 × 112	7 × 7, 64, stride 2				
		3 × 3 max pool, stride 2				
conv2-x	56 × 56	3 × 3, 64 3 × 3, 64 × 2	3 × 3, 64 3 × 3, 64 × 3	1 × 1, 64 3 × 3, 64 × 3 1 × 1, 256	1 × 1, 64 3 × 3, 64 × 3 1 × 1, 256	1 × 1, 64 3 × 3, 64 × 3 1 × 1, 256
conv3-x	28 × 28	3 × 3, 128 3 × 3, 128 × 2	3 × 3, 128 3 × 3, 128 × 4	1 × 1, 128 3 × 3, 128 × 4 1 × 1, 512	1 × 1, 128 3 × 3, 128 × 4 1 × 1, 512	1 × 1, 128 3 × 3, 128 × 8 1 × 1, 512
conv4-x	14 × 14	3 × 3, 256 3 × 3, 256 × 2	3 × 3, 256 3 × 3, 256 × 6	1 × 1, 256 3 × 3, 256 × 6 1 × 1, 1024	1 × 1, 256 3 × 3, 256 × 23 1 × 1, 1024	1 × 1, 256 3 × 3, 256 × 36 1 × 1, 1024
conv5-x	7 × 7	3 × 3, 512 3 × 3, 512 × 2	3 × 3, 512 3 × 3, 512 × 3	1 × 1, 512 3 × 3, 512 × 3 1 × 1, 2048	1 × 1, 512 3 × 3, 512 × 3 1 × 1, 2048	1 × 1, 512 3 × 3, 512 × 3 1 × 1, 2048
	1 × 1	average pool, 1000-d fc, softmax				
FLOPs		1.8 × 10 ⁹	3.6 × 10 ⁹	3.8 × 10 ⁹	7.6 × 10 ⁹	11.3 × 10 ⁹

Table 1. Architectures for ImageNet. Building blocks are shown in brackets (see also Fig. 5), with the numbers of blocks stacked. Down-sampling is performed by conv3-1, conv4-1, and conv5-1 with a stride of 2.

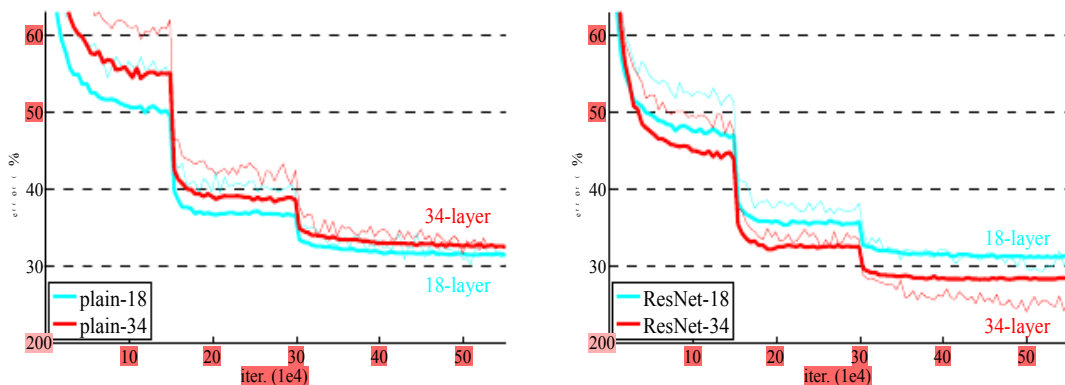


Figure 4. Training on ImageNet. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	25.03

Table 2. Top-1 error (%) 10-crop testing) on ImageNet validation. Here the ResNets have no extra parameter compared to their plain counterparts. Fig. 4 shows the training procedures.

34-layer plain net has higher training error throughout the whole training procedure, even though the solution space of the 18-layer plain network is a subspace of that of the 34-layer one.

We argue that this optimization difficulty is unlikely to be caused by vanishing gradients. These plain networks are trained with BN [16], 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. In fact, the 34-layer plain net is still able to achieve competitive accuracy (Table 3), suggesting that the solver works to some extent. We conjecture that the deep plain nets may have exponentially low convergence rates, which impact the

reducing of the training error³. The reason for such optimization difficulties will be studied in the future.

Residual Networks. Next we evaluate 18-layer and 34-layer residual nets (ResNets). The baseline architectures are the same as the above plain nets, except that a shortcut connection is added to each pair of 3 × 3 filters as in Fig. 3 (right). In the first comparison (Table 2 and Fig. 4 right), we use identity mapping for all shortcuts and zero-padding for increasing dimensions (option A). So they have no extra parameter compared to the plain counterparts.

We have three major observations from Table 2 and Fig. 4. First, the situation is reversed with residual learning – the 34-layer ResNet is better than the 18-layer ResNet (by 2.8%). More importantly, the 34-layer ResNet exhibits considerably lower training error and is generalizable to the validation data. This indicates that the degradation problem is well addressed in this setting and we manage to obtain accuracy gains from increased depth.

Second, compared to its plain counterpart, the 34-layer

³We have experimented with more training iterations (3×) and still observed the degradation problem, suggesting that this problem cannot be feasibly addressed by simply using more iterations.

model	top-1 err.	top-5 err.
VGG-16 [40]	28.07	9.33
GoogLeNet [43]	-	9.15
PReLU-net [12]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	21.43	5.71

Table 3. Error rates (%; 10-crop testing) on ImageNet validation. VGG-16 is based on our test. ResNet-50/101/152 are of option B that only uses projections for increasing dimensions.

method	top-1 err.	top-5 err.
VGG [40] (ILSVRC'14)	-	8.43 [†]
GoogLeNet [43] (ILSVRC'14)	-	7.89
VGG [40] (v5)	24.4	7.1
PReLU-net [12]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

Table 4. Error rates (%) of single-model results on the ImageNet validation set (except [†] reported on the test set).

method	top-5 err. (test)
VGG [40] (ILSVRC'14)	7.32
GoogLeNet [43] (ILSVRC'14)	6.66
VGG [40] (v5)	6.8
PReLU-net [12]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	3.57

Table 5. Error rates (%) of ensembles. The top-5 error is on the test set of ImageNet and reported by the test server.

ResNet reduces the top-1 error by 3.5% (Table 2), resulting from the successfully reduced training error (Fig. 4 right vs. left). This comparison verifies the effectiveness of residual learning on extremely deep systems.

Last, we also note that the 18-layer plain/residual nets are comparably accurate (Table 2), but the 18-layer ResNet converges faster (Fig. 4 right vs. left). When the net is “not overly deep” (18 layers here), the current SGD solver is still able to find good solutions to the plain net. In this case, the ResNet eases the optimization by providing faster convergence at the early stage.

Identity vs. Projection Shortcuts. We have shown that

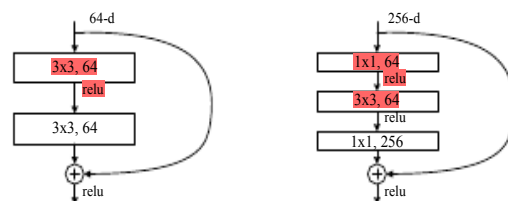


Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a “bottleneck” building block for ResNet-50/101/152.

parameter-free, identity shortcuts help with training. Next we investigate projection shortcuts (Eqn.(2)). In Table 3 we compare three options: (A) zero-padding shortcuts are used for increasing dimensions, and all shortcuts are parameter-free (the same as Table 2 and Fig. 4 right); (B) projection shortcuts are used for increasing dimensions, and other shortcuts are identity; and (C) all shortcuts are projections.

Table 3 shows that all three options are considerably better than the plain counterpart. B is slightly better than A. We argue that this is because the zero-padded dimensions in A indeed have no residual learning. C is marginally better than B, and we attribute this to the extra parameters introduced by many (thirteen) projection shortcuts. But the small differences among A/B/C indicate that projection shortcuts are not essential for addressing the degradation problem. So we do not use option C in the rest of this paper, to reduce memory/time complexity and model sizes. Identity shortcuts are particularly important for not increasing the complexity of the bottleneck architectures that are introduced below.

Deeper Bottleneck Architectures. Next we describe our deeper nets for ImageNet. Because of concerns on the training time that we can afford, we modify the building block as a bottleneck design⁴. For each residual function $\mathcal{F}$, we use a stack of 3 layers instead of 2 (Fig. 5). The three layers are 1×1 , 3×3 , and 1×1 convolutions, where the 1×1 layers are responsible for reducing and then increasing (restoring) dimensions, leaving the 3×3 layer a bottleneck with smaller input/output dimensions. Fig. 5 shows an example, where both designs have similar time complexity.

The parameter-free identity shortcuts are particularly important for the bottleneck architectures. If the identity shortcut in Fig. 5 (right) is replaced with projection, one can show that the time complexity and model size are doubled, as the shortcut is connected to the two high-dimensional ends. So identity shortcuts lead to more efficient models for the bottleneck designs.

50-layer ResNet: We replace each 2-layer block in the

⁴Deeper non-bottleneck ResNets (e.g., Fig. 5 left) also gain accuracy from increased depth (as shown on CIFAR-10), but are not as economical as the bottleneck ResNets. So the usage of bottleneck designs is mainly due to practical considerations. We further note that the degradation problem of plain nets is also witnessed for the bottleneck designs.

34-layer net with this 3-layer bottleneck block, resulting in a 50-layer ResNet (Table 1). We use option B for increasing dimensions. This model has 3.8 billion FLOPs.

101-layer and 152-layer ResNets: We construct 101-layer and 152-layer ResNets by using more 3-layer blocks (Table 1). Remarkably, although the depth is significantly increased, the 152-layer ResNet (11.3 billion FLOPs) still has lower complexity than VGG-16/19 nets (15.3/19.6 billion FLOPs).

The 50/101/152-layer ResNets are more accurate than the 34-layer ones by considerable margins (Table 3 and 4). We do not observe the degradation problem and thus enjoy significant accuracy gains from considerably increased depth. The benefits of depth are witnessed for all evaluation metrics (Table 3 and 4).

Comparisons with State-of-the-art Methods. In Table 4 we compare with the previous best single-model results. Our baseline 34-layer ResNets have achieved very competitive accuracy. Our 152-layer ResNet has a single-model top-5 validation error of 4.49%. This single-model result outperforms all previous ensemble results (Table 5). We combine six models of different depth to form an ensemble (only with two 152-layer ones at the time of submitting). This leads to 3.57% top-5 error on the test set (Table 5). This entry won the 1st place in ILSVRC 2015.

4.2. CIFAR-10 and Analysis

We conducted more studies on the CIFAR-10 dataset [20], which consists of 50k training images and 10k testing images in 10 classes. We present experiments trained on the training set and evaluated on the test set. Our focus is on the behaviors of extremely deep networks, but not on pushing the state-of-the-art results, so we intentionally use simple architectures as follows.

The plain/residual architectures follow the form in Fig. 3 (middle/right). The network inputs are 32×32 images, with the per-pixel mean subtracted. The first layer is 3×3 convolutions. Then we use a stack of 6^n layers with 3×3 convolutions on the feature maps of sizes $\langle 32, 16, 8 \rangle$ respectively, with 2^n layers for each feature map size. The numbers of filters are $\langle 16, 32, 64 \rangle$ respectively. The subsampling is performed by convolutions with a stride of 2. The network ends with a global average pooling, a 10-way fully-connected layer, and softmax. There are totally $6^n + 2$ stacked weighted layers. The following table summarizes the architecture:

output map size	32×32	16×16	8×8
# layers	$1 + 2^n$	2^n	2^n
# filters	16	32	64

When shortcut connections are used, they are connected to the pairs of 3×3 layers (totally 3^n shortcuts). On this dataset we use identity shortcuts in all cases (i.e., option A),

method		error (%)	
Maxout [9]		9.38	
NIN [25]		8.81	
DSN [24]		8.22	
	# layers	# params	
FitNet [34]	19	2.5M	8.39
Highway [41, 42]	19	2.3M	7.54 (7.72± 0.16)
Highway [41, 42]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	6.43 (6.61± 0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the CIFAR-10 test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show “best (mean $\pm$ std)” as in [42].

so our residual models have exactly the same depth, width, and number of parameters as the plain counterparts.

We use a weight decay of 0.0001 and momentum of 0.9, and adopt the weight initialization in [12] and BN [16] but with no dropout. These models are trained with a mini-batch size of 128 on two GPUs. We start with a learning rate of 0.1, divide it by 10 at 32k and 48k iterations, and terminate training at 64k iterations, which is determined on a 45k/5k train/val split. We follow the simple data augmentation in [24] for training: 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded image or its horizontal flip. For testing, we only evaluate the single view of the original 32×32 image.

We compare $n = \langle 3, 5, 7, 9 \rangle$ leading to 20, 32, 44, and 56-layer networks. Fig. 6 (left) shows the behaviors of the plain nets. The deep plain nets suffer from increased depth, and exhibit higher training error when going deeper. This phenomenon is similar to that on ImageNet (Fig. 4, left) and on MNIST (see [41]), suggesting that such an optimization difficulty is a fundamental problem.

Fig. 6 (middle) shows the behaviors of ResNets. Also similar to the ImageNet cases (Fig. 4, right), our ResNets manage to overcome the optimization difficulty and demonstrate accuracy gains when the depth increases.

We further explore $n = 18$ that leads to a 110-layer ResNet. In this case, we find that the initial learning rate of 0.1 is slightly too large to start converging⁵. So we use 0.01 to warm up the training until the training error is below 80% (about 400 iterations), and then go back to 0.1 and continue training. The rest of the learning schedule is as done previously. This 110-layer network converges well (Fig. 6, middle). It has fewer parameters than other deep and thin

⁵With an initial learning rate of 0.1, it starts converging ($\leq 90\%$ error) after several epochs, but still reaches similar accuracy.

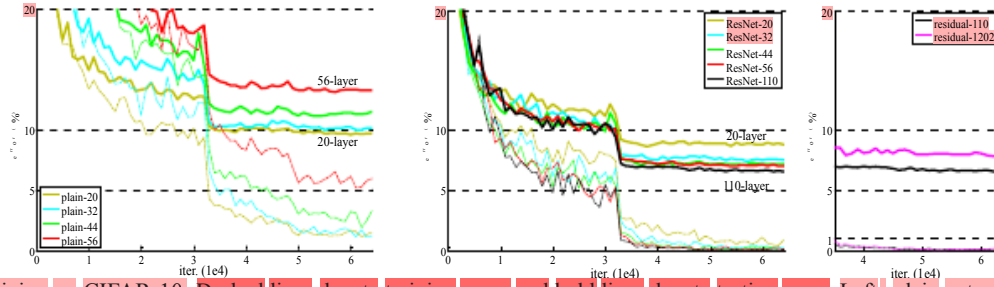


Figure 6. Training on CIFAR-10. Dashed lines denote training error, and bold lines denote testing error. Left: plain networks. The error of plain-110 is higher than 60% and not displayed. Middle: ResNets. Right: ResNets with 110 and 1202 layers.

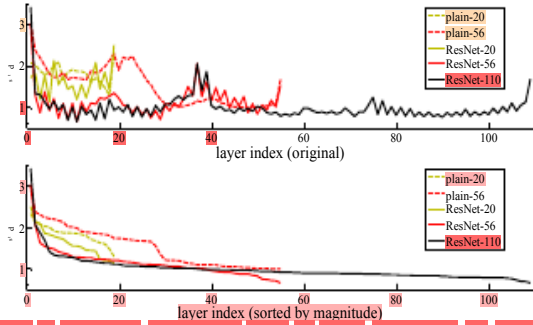


Figure 7. Standard deviations (std) of layer responses on CIFAR-10. The responses are the outputs of each 3×3 layer, after BN and before nonlinearity. Top: the layers are shown in their original order. Bottom: the responses are ranked in descending order.

networks such as FitNet [34] and Highway [41] (Table 6), yet is among the state-of-the-art results (6.43%, Table 6).

Analysis of Layer Responses. Fig. 7 shows the standard deviations (std) of the layer responses. The responses are the outputs of each 3×3 layer, after BN and before other nonlinearity (ReLU/addition). For ResNets, this analysis reveals the response strength of the residual functions. Fig. 7 shows that ResNets have generally smaller responses than their plain counterparts. These results support our basic motivation (Sec.3.1) that the residual functions might be generally closer to zero than the non-residual functions. We also notice that the deeper ResNet has smaller magnitudes of responses, as evidenced by the comparisons among ResNet-20, 56, and 110 in Fig. 7. When there are more layers, an individual layer of ResNets tends to modify the signal less.

Exploring Over 1000 layers. We explore an aggressively deep model of over 1000 layers. We set $n = 200$ that leads to a 1202-layer network, which is trained as described above. Our method shows no optimization difficulty, and this 10^3 -layer network is able to achieve training error $< 0.1\%$ (Fig. 6, right). Its test error is still fairly good (7.93%, Table 6).

But there are still open problems on such aggressively deep models. The testing result of this 1202-layer network is worse than that of our 110-layer network, although both

training data	07+12	07++12
test data	VOC 07 test	VOC 12 test
VGG-16	73.2	70.4
ResNet-101	76.4	73.8

Table 7. Object detection mAP (%) on the PASCAL VOC 2007/2012 test sets using baseline Faster R-CNN. See also appendix for better results.

metric	mAP@.5	mAP@[.5, .95]
VGG-16	41.5	21.2
ResNet-101	48.4	27.2

Table 8. Object detection mAP (%) on the COCO validation set using baseline Faster R-CNN. See also appendix for better results.

have similar training error. We argue that this is because of overfitting. The 1202-layer network may be unnecessarily large (19.4M) for this small dataset. Strong regularization such as maxout [9] or dropout [13] is applied to obtain the best results ([9, 25, 24, 34]) on this dataset. In this paper, we use no maxout/dropout and just simply impose regularization via deep and thin architectures by design, without distracting from the focus on the difficulties of optimization. But combining with stronger regularization may improve results, which we will study in the future.

4.3. Object Detection on PASCAL and MS COCO

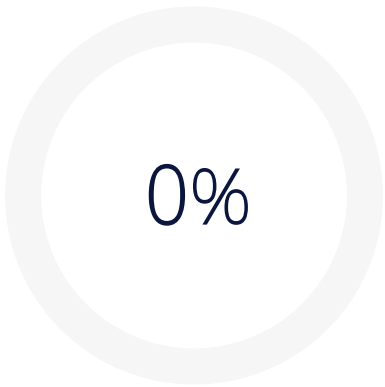
Our method has good generalization performance on other recognition tasks. Table 7 and 8 show the object detection baseline results on PASCAL VOC 2007 and 2012 [5] and COCO [26]. We adopt Faster R-CNN [32] as the detection method. Here we are interested in the improvements of replacing VGG-16 [40] with ResNet-101. The detection implementation (see appendix) of using both models is the same, so the gains can only be attributed to better networks. Most remarkably, on the challenging COCO dataset we obtain a 6.0% increase in COCO's standard metric (mAP@[.5, .95]), which is a 28% relative improvement. This gain is solely due to the learned representations.

Based on deep residual nets, we won the 1st places in several tracks in ILSVRC & COCO 2015 competitions: ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation. The details are in the appendix.

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] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.
- [3] W. L. Briggs, S. F. McCormick, et al. *A Multigrid Tutorial*. Siam, 2000.
- [4] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.
- [5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, pages 303–338, 2010.
- [6] R. Girshick. Fast R-CNN. In *ICCV*, 2015.
- [7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [8] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.
- [9] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. arXiv:1302.4389, 2013.
- [10] K. He and J. Sun. Convolutional neural networks at constrained time cost. In *CVPR*, 2015.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.
- [13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.
- [14] S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. Diploma thesis, TU Munich, 1991.
- [15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [17] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*, 33, 2011.
- [18] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 2012.
- [19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093, 2014.
- [20] A. Krizhevsky. Learning multiple layers of features from tiny images. Tech Report, 2009.
- [21] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989.
- [23] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–50. Springer, 1998.
- [24] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. arXiv:1409.5185, 2014.
- [25] M. Lin, Q. Chen, and S. Yan. Network in network. arXiv:1312.4400, 2013.
- [26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.
- [27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [28] G. Montu'far, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *NIPS*, 2014.
- [29] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.
- [30] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
- [31] T. Raiko, H. Valpola, and Y. LeCun. Deep learning made easier by linear transformations in perceptrons. In *AISTATS*, 2012.
- [32] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [33] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 1996.
- [34] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. arXiv:1409.0575, 2014.
- [36] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2013.
- [37] N. N. Schraudolph. Accelerated gradient descent by factor-centering decomposition. Technical report, 1998.
- [38] N. N. Schraudolph. Centering neural network gradient factors. In *Neural Networks: Tricks of the Trade*, pages 207–226. Springer, 1998.
- [39] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.
- [40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [41] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. arXiv:1505.00387, 2015.
- [42] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. 1507.06228, 2015.
- [43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [44] R. Szeliski. Fast surface interpolation using hierarchical basis functions. *TPAMI*, 1990.
- [45] R. Szeliski. Locally adapted hierarchical basis preconditioning. In *SIGGRAPH*, 2006.
- [46] T. Vatanen, T. Raiko, H. Valpola, and Y. LeCun. Pushing stochastic gradient towards second-order methods—backpropagation learning with transformations in nonlinearities. In *Neural Information Processing*, 2013.
- [47] A. Vedaldi and B. Fulkerson. *VLFeat: An open and portable library of computer vision algorithms*, 2008.
- [48] W. Venables and B. Ripley. *Modern applied statistics with s-plus*. 1999.
- [49] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In *ECCV*, 2014.

AI Content



	Text coverage	Words
☒ AI Text	0%	0
☐ Human Text	100%	7,627
Excluded		
☐ Omitted Words		0

About AI Detection

Our AI Detector is the only enterprise-level solution that can verify if the content was written by a human or generated by AI, including source code and text that has been plagiarized or modified. [Learn more](#)

☒ AI Text

A body of text that has been generated or altered by AI technology. [Learn more](#)

☐ Human Text

Any text that has been fully written by a human and has not been altered or generated by AI. [Learn more](#)

Copyleaks AI Detector Effectiveness

Credible data at scale, coupled with machine learning and widespread adoption, allows us to continually refine and improve our ability to understand complex text patterns, resulting in over 99% accuracy—far higher than any other AI detector—and improving daily. [Learn more](#)

Ideal Text Length

The higher the character count, the easier for our technology to determine irregular patterns, which results in a higher confidence rating for AI detection. [Learn more](#)

Reasons It Might Be AI When You Think It's Not

The AI Detector can detect a variety of AI-generated text, including tools that use AI technology to paraphrase content, auto-complete sentences, and more. [Learn more](#)

User AI Alert History

Historical data of how many times a user has been flagged for potentially having AI text within their content. [Learn more](#)

AI Logic

The number of times a phrase was found more frequently in AI vs human text is shown according to low, medium, and high frequency. [Learn more](#)