Structured Review: Deep Residual Learning for Image Recognition (1512.03385v1)

Synopsis of the paper

The paper addresses the degradation problem in training very deep convolutional networks: as depth increases, accuracy saturates and then degrades, and this is not due to overfitting (see Fig. 1; Sec. 1). The authors propose a residual learning framework in which stacked layers learn residual functions F(x) := H(x) − x with reference to layer inputs, so that the overall mapping is F(x) + x, implemented via identity shortcut connections (Eq. 1; Fig. 2). They provide comprehensive experiments on ImageNet and CIFAR-10 showing that residual networks (ResNets) are easier to optimize and gain accuracy from increased depth (e.g., 34-layer ResNet vs 34-layer plain in Table 2 and Fig. 4; 152-layer ResNet on ImageNet). Reported results include 3.57% top-5 error on the ImageNet test set (ensemble), ILSVRC 2015 classification winner, and strong gains on COCO detection and other tasks (Sec. 1; Tables 3–5, 7–8). The paper also explores networks with over 1000 layers on CIFAR-10 (Sec. 4.2; Table 6).

Summary of Review

The paper presents a clear, well-motivated solution to the degradation problem in deep nets, with strong empirical support on ImageNet and CIFAR-10 and convincing comparisons between plain and residual nets of equal depth/parameters (see Table 2; Fig. 4; Sec. 4.1). The formulation is simple (Eq. 1–2), implementation details and architectures are described (Sec. 3.3–3.4; Table 1), and the extension to detection/localization is demonstrated (Sec. 4.3; Tables 7–8, 10–14). Limitations include: no theoretical analysis of why residual mappings are easier to optimize (the hypothesis in Sec. 3.1 is acknowledged as open; footnote 2); the 1202-layer CIFAR-10 model generalizes worse than the 110-layer one (Table 6; Sec. 4.2); and some notation and derivations could be tightened (e.g., Eq. 1 and the role of σ in Fig. 2). Overall the work is impactful and reproducible, with evidence-backed claims and appropriate discussion of alternatives and future work.

Strengths

Clarity of problem and motivation

- The degradation problem is stated precisely and illustrated with plain 20-layer vs 56-layer training/test error (Fig. 1) and with ImageNet plain-18 vs plain-34 (Table 2; Fig. 4 left), showing that deeper plain nets have higher training error (Sec. 1; Sec. 4.1).
- The residual reformulation $H(x) \rightarrow F(x) + x$ is motivated by the existence of a constructed solution (identity for added layers) and the difficulty of learning identity with stacked nonlinear layers (Sec. 1; Sec. 3.1).
- Shortcut connections are described concretely (identity vs projection, option A/B/C) with minimal extra parameters and complexity (Eq. 1–2; Fig. 2–3; Sec. 3.2–3.3).

Rigorous and controlled experimentation

- Plain vs residual nets are compared under matched depth, width, and parameter count (e.g., ResNet-34 A has no extra parameters; Table 2; Fig. 4 right; Sec. 4.1), isolating the effect of residual learning.
- Multiple depths (18/34 on ImageNet; 20/32/44/56/110/1202 on CIFAR-10) and shortcut options (A, B, C) are evaluated (Tables 2–3, 6; Fig. 4–6), with training curves and layer-response analysis (Fig. 7) supporting the “small residual” motivation (Sec. 3.1; Sec. 4.2).

- Implementation details are given (data augmentation, BN, learning rate schedule, FLOPs; Sec. 3.4; Table 1), facilitating reproduction.

Generality and impact

- Results transfer to object detection and localization (Faster R-CNN with ResNet-101; Tables 7–8, 10–14; Appendix A–C), with large relative gains (e.g., 28% on COCO mAP@[.5,.95]; Sec. 4.3).
- Very deep ResNets (152-layer on ImageNet, 110-layer on CIFAR-10) are trained successfully without degradation (Tables 3–4, 6), and the 1000+ layer exploration (1202-layer) is reported honestly, including worse generalization than 110-layer (Table 6; Sec. 4.2).

Related work and writing

- Related work on residual representations, shortcut connections, and highway networks is discussed, with clear differentiation (e.g., identity vs gated shortcuts; Sec. 2).
- The paper is well structured (problem → method → experiments → detection/localization) and uses consistent notation and figure/table references.

Weaknesses

Mathematical formulation and notation

- In Eq. (1), y = F(x, {Wi}) + x, the text states “The operation F + x is performed by a shortcut connection and element-wise addition. We adopt the second nonlinearity after the addition (i.e., σ(y), see Fig. 2)” (Sec. 3.2). Fig. 2 shows a path x → F(x) → F(x)+x → σ(y), but Eq. (1) does not explicitly include σ(y) as the output of the block; the relationship between “y” in Eq. (1) and the post-σ output could be stated more explicitly to avoid ambiguity.
- For the two-layer block, “F = W2σ(W1x)” is given with “biases are omitted for simplifying notations” (Sec. 3.2). The convention (e.g., whether bias is absorbed into W or omitted in the analysis) is not restated where needed, and the step from this F to the full block output σ(F(x)+x) is left implicit.
- When dimensions change, Eq. (2) uses y = F(x, {Wi}) + Ws x with Ws for projection; the text says “Ws is only used when matching dimensions” (Sec. 3.2). The indexing of which shortcut uses Ws (e.g., “dotted” in Fig. 3) is clear from the

figure, but a single sentence tying Eq. (2) to “option B” in the same subsection would improve consistency.
- The hypothesis that “multiple nonlinear layers can asymptotically approximate complicated functions” is cited as an open question (footnote 2; Sec. 3.1). The logical step from “approximating H(x)” to “approximating H(x)−x” is correct, but the benefit of the reformulation is argued heuristically (“ease of learning might be different”) without a formal statement of what “easier” means (e.g., optimization landscape, convergence rate).

Limited theoretical grounding

- The claim that optimizing the residual mapping is easier than the original mapping is supported empirically (e.g., small layer responses in Fig. 7; Sec. 4.2) but not by convergence or landscape analysis (Sec. 3.1). The conjecture on “exponentially low convergence rates” for deep plain nets (Sec. 4.1, footnote 3) is not formalized.
- No direct comparison of gradient norms or loss curvature between plain and residual nets is given; the “healthy norms with BN” check (Sec. 4.1) only concerns plain nets.

Overfitting and extreme depth

- The 1202-layer ResNet on CIFAR-10 has higher test error (7.93%) than the 110-layer (6.43%) despite similar training behavior (Table 6; Fig. 6 right; Sec. 4.2). The authors attribute this to overfitting and dataset size; no ablation (e.g., stronger regularization or different depth schedules) is reported for the 1202-layer case.
- CIFAR-10 uses “no maxout/dropout” (Sec. 4.2); the choice is justified by focus on optimization, but the impact on the 1202-layer result is not quantified.

Suggestions for Improvement

Mathematical formulation and notation

- Add one sentence after Eq. (1) stating explicitly that the block’s final output is σ(y) (or introduce a symbol for the post-σ output) so that Eq. (1) and Fig. 2 are aligned unambiguously (Sec. 3.2).
- In the two-layer example “F = W2σ(W1x)”, briefly state whether the argument of the outer σ in the full block is F(x)+x, and confirm that “y” in Eq. (1) is the pre-σ value; optionally add a single line: “block output = σ(F(x)+x).”

- In Sec. 3.2, add a short phrase linking “projection shortcut” and “option B” (and dimension-increasing shortcuts in Fig. 3) to Eq. (2) so that readers can match equation and implementation in one pass.
- In Sec. 3.1, add a sentence that clarifies “easier to optimize” (e.g., in terms of optimization landscape, effective conditioning, or convergence speed) even if only at an intuitive level, and keep the open hypothesis in footnote 2.

Limited theoretical grounding

- Consider adding a small subsection or paragraph summarizing gradient-norm or loss-surface observations (if available) for plain vs residual nets of the same depth, to support the “easier optimization” claim; if such experiments are not feasible, state so explicitly.
- Optionally cite or briefly discuss existing work on loss landscape or convergence of deep nets (e.g., in relation to [28] or initialization [13, 23]) to position the residual reformulation within the theory of optimization.

Overfitting and extreme depth

- For the 1202-layer CIFAR-10 model, add one or two ablations: e.g., (1) applying dropout or stronger weight decay and reporting test error, or (2) reporting validation error vs iterations to show overfitting onset. This would make the overfitting explanation evidence-based.
- In Sec. 4.2, add one sentence on whether the same training protocol (e.g., learning rate warmup) was used for 110- and 1202-layer nets and whether any extra regularization was tried for the 1202-layer net; this would clarify reproducibility and future work.

References

[Bengio et al., 1994] Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994.

[Bishop, 1995] Bishop, C. M. Neural networks for pattern recognition. Oxford University Press, 1995.

[Briggs et al., 2000] Briggs, W. L., McCormick, S. F., et al. A Multigrid Tutorial. SIAM, 2000.

[Glorot & Bengio, 2010] Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.

[He & Sun, 2015] He, K. and Sun, J. Convolutional neural networks at constrained time cost. In CVPR, 2015.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In ICCV, 2015.

[Ioffe & Szegedy, 2015] Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[Montúfar et al., 2014] Montúfar, G., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In NIPS, 2014.

[Nair & Hinton, 2010] Nair, V. and Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In ICML, 2010.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[Russakovsky et al., 2014] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, M., Bernstein, M., et al. ImageNet large scale visual recognition challenge. arXiv:1409.0575, 2014.

[Simonyan & Zisserman, 2015] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[Srivastava et al., 2015] Srivastava, R. K., Greff, K., and Schmidhuber, J. Highway networks. arXiv:1505.00387, 2015.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, V., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In CVPR, 2015.