Paper Review:"Are ResNets Provably Better than Linear Predictors? "

Author: Ohad Shamir
Reviewer: Orhun Güley

Technical University of Munich - Master Seminar Theoretical advances in Deep Learning

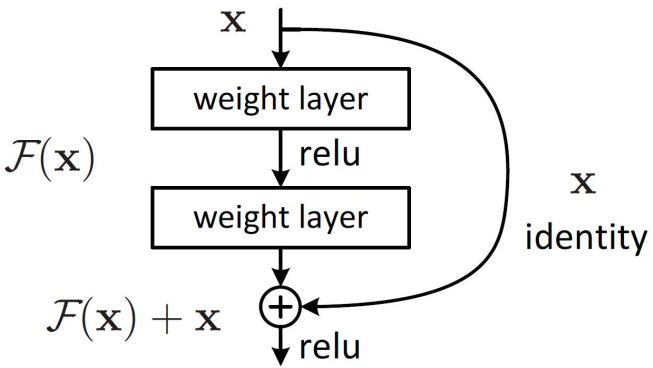
25.06.2020

What is Residual Networks?

- Consist of "residual units" which has a mathematical form of $\mathbf{y} = f\left(h(\mathbf{x}) + g_{\Phi}(\mathbf{x})\right)$, where f and h are fixed functions.
- In most of the case f and h are just idenity and the residual units turn into the form $\mathbf{y} = \mathbf{x} + g_{\Phi}(\mathbf{x})$
- Allows network to focus on the ''residual" of the network.

Success of ResNets

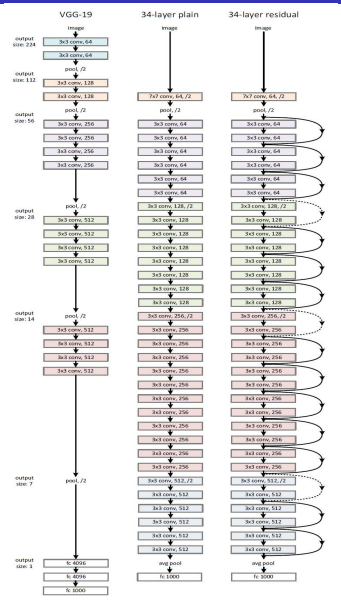
- "Shortcut connection" property



Success of ResNets

- Emprically shown to be very successful.

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**



Goal of the Paper

- Rigorously analyzing the performance of ResNets compared to linear predictors.
- Limited theoretical understanding on training of deep learning models.
- Most of existing rigorous theoretical results a limited to linear networks, which are not using non-linear activation functions.

Past Work

- There is also similar works but not the same with ResNet architecture, which the results provably do not hold on ResNets.
  - $\mathbf{x} \mapsto f_S(\mathbf{x}) + f_D(\mathbf{x})$, where f_S is a one-hidden-layer network, and f_D is an arbitrary, possibly deeper network.
  - Unlike ResNets, they don't have final tunable layer to combine the outputs of f_S and f_D.
  - Assumptions are non-trivial, do not apply as-is to standard activations and losses like ReLU and logistic loss
  - Requires specific conditions on the data, such as linear separability or a certain low-rank structure.

- Set $g_\Phi(\mathbf{x}) = Vf_\theta(\mathbf{x})$, and the corresponding residual network outputs

$$\mathbf{x} \mapsto \mathbf{w}^\top (\mathbf{x} + Vf_\theta(\mathbf{x}))$$

- Assuming that the network is trained with respect to some data distribution, using a loss function $\ell(p, y)$, where p is prediction and y is the target value. Then we have the following optimization problem.

$$\min_{\mathbf{w}, V, \theta} F(\mathbf{w}, V, \theta) := \mathbb{E}_{\mathbf{x}, y} \left[\ell \left(\mathbf{w}^\top (\mathbf{x} + Vf_\theta(\mathbf{x})) ; y \right) \right]$$

- **Assumption 1.** For any y, the loss $\ell(p, y)$ is twice differentiable w.r.t (w, V) and convex in p. Note that it doesnt have to be differentiable w.r.t θ
- For linear predictior, the following notation is used;

$$F_{\text{lin}}(\mathbf{w}) := F(\mathbf{w}, \mathbf{0}, \theta) = \mathbb{E}_{\mathbf{x},y}\left[\ell\left(\mathbf{w}^\top \mathbf{x}; y\right)\right]$$

Definition 1. ($\epsilon - \boldsymbol{SOPSP}$).

- Stands for ϵ-second order partial stationary point, where second order partial derivative(gradient) is ϵ.
- It is trivial that any local minimum of the point (w, V, θ) of F is 0-SOPSP since the gradient of F with respect to (w, V) will be zero(since the norm of the gradient will be zero as well).

Competitiveness with Linear Predictors

- Theorem 3 and Corollary 1 are the main results of the paper, which are proven in 2 stages.
- In the first stage - Theorem 1 -, for any point such that $\mathbf{w} \neq 0$, the difference between F and $F_{lin}(w^*)$ is upper-bounded with some term dependent on $\|\nabla F_\theta(\mathbf{w}, V)\|$.
- In the second stage - Theorem 2 -, the case $\mathbf{w} = 0$ is considered and objective value of F is upper-bounded with objective value of F_{lin} plus some value $\|\nabla F_\theta(\mathbf{w}, V)\|$.

Theorem 1. At any point $(\mathbf{w}, V, \theta)$ such that $\mathbf{w} \neq 0$, and for any vector $\mathbf{w}^*$ of the same dimension as $\mathbf{w}$

$$\|\nabla F_\theta(\mathbf{w}, V)\| \geq \frac{F(\mathbf{w}, V, \theta) - F_{\text{lin}}(\mathbf{w}^*)}{\sqrt{2\|\mathbf{w}\|^2 + \|\mathbf{w}^*\|^2 \left(2 + \frac{\|V\|^2}{\|\mathbf{w}\|^2}\right)}}$$

- Implies that for any point $(\mathbf{w}, V, \theta)$ such that $\mathbf{w} \neq 0$, if objective value of residual net is larger than the objective value of linear predictor, $\|\nabla F_\theta(\mathbf{w}, V)\|$ is non-zero, which means that point cannot be a stationary point or a local minimum of F.

Theorem 2. For any V, θ, w^*,

$$\lambda_{\min}\left(\nabla^2 F_\theta(0, V)\right) \leq 0 \qquad \textit{and}$$

$$\frac{\|\nabla F_\theta(0, V)\| + }{\|V\|\sqrt{|\lambda_{\min}\left(\nabla^2 F_\theta(0, V)\right)| \cdot \left\|\frac{\partial^2}{\partial \mathbf{w}^2} F_\theta(0, V)\right\| + \lambda_{\min}\left(\nabla^2 F_\theta(0, V)\right)^2}}$$

$$\geq \frac{F(0, V, \theta) - F_{lin}(\mathrm{w}^*)}{\|\mathrm{w}^*\|}$$

- The case $w = 0$ is considered and objective value of F is upper-bounded with optimal objective value of F_{lin} plus a term.

Competitiveness with Linear Predictors

In **Theorem 3**, paper cobines the results from **Theorem 1** and **Theorem 2**, and proposes a general inequality under the following assumptions;

- $\max\{\|\mathbf{w}\|, \|V\|\} \leq b$
- $F_\theta(\mathbf{w}, V), \nabla F_\theta(\mathbf{w}, V)$ and $\nabla^2 F_\theta(\mathbf{w}, V)$ are μ_0-Lipschitz, μ_1-Lipschitz, and μ_2-Lipschitz in $(\mathbf{w}, V)$ respectively.
- any $(\mathbf{w}, V, \theta) \in \mathcal{W}$, we have $(\mathbf{0}, V, \theta) \in \mathcal{W}$ and $\left\|\nabla^2 F_\theta(\mathbf{0}, V)\right\| \leq \mu_1$ where b, r, μ_0, μ_1, μ_2 are fixed positive numbers and $\epsilon \geq 0$

Competitiveness with Linear Predictors

Then for any $(\mathbf{w}, V, \theta) \in \mathcal{M}$ which is an $\epsilon-$SOPSP of F on $\mathcal{M}$

$$F(\mathbf{w}, V, \theta) \leq \min_{\mathbf{w}:\|\mathbf{w}\|\leq r} F_{lin}(\mathbf{w}) + (\epsilon + \sqrt[4]{\epsilon}) \cdot \text{poly}\,(b, r, \mu_0, \mu_1, \mu_2)$$

- Any local minima of function F is 0-SOPSP.
- It is trivial to see that in the case of $\epsilon = 0$, the polynomial term in **Theorem 3** will be extracted from the inequality and we get the following **corollary**:

$F(\mathbf{w}, V, \theta) \leq \inf_{\mathbf{w}} F_{lin}(\mathbf{w})$, for every local minimum of function F

Effects of Norm and Regularization

The bound at the Theorem 3 highly depend on the terms $\|\mathbf{w}\|$ and $\|V\|$

Example 1. Fix some $\epsilon > 0$. Suppose $\mathbf{x}, \mathbf{w}, V, \mathbf{w}^*$ are all scalars, $\mathbf{w}^* = 1, f_\theta(\mathbf{x}) = \epsilon\mathbf{x}$ (with no dependence on a parameter θ), $\ell(p; y) = \frac{1}{2}(p - y)^2$ is the squared loss, and $\mathbf{x} = y = 1$ w.p. 1. Then the objective can be equivalently written as

$$F(w, v) = \frac{1}{2}(w(1 + \epsilon v) - 1)^2$$

Then the gradient and Hessian of $F(w, v)$ equal

$$\begin{pmatrix} (w - 1 + \epsilon wv)(1 + \epsilon v) \\ (w - 1 + \epsilon wv)\epsilon w \end{pmatrix} \text{ and}$$

$$\begin{pmatrix} (1 + \epsilon v)^2 & \epsilon(2w + 2\epsilon wv - 1) \\ \epsilon(2w + 2\epsilon wv - 1) & \epsilon^2 w^2 \end{pmatrix}$$

The gradient is 0 at $(w, v) = (0, -1/\epsilon)$,, and the Hessian equals $\begin{pmatrix} 0 & -\epsilon \\ -\epsilon & 0 \end{pmatrix}$ which is arbitrarily close to 0 if ϵ is small enough.

However, the objective value at that point is equal to

$$F\left(0, -\frac{1}{\epsilon}\right) = \frac{1}{2}$$

$$F_{lin}(1) = 0$$

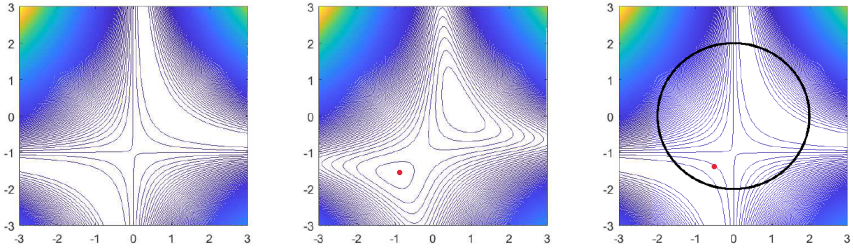
$$F\left(0, -\frac{1}{\epsilon}\right) > 0 = F_{lin}(1)$$

- Implies that norm of the arguments have an important role in the theorem!

Effects of Norm and Regularization

- Adding a regularization term is also considered in order to keep the norm of weighs low.
- However, this might cause adding new spurious local minimas that did not exist in $F(\mathbf{w}, V, \theta)$.

Effects of Norm and Regularization



-contour plots:

a) without regularization term
b) regularization term added to objective function
c) norm of w and v is constrained

- x-axis corresponds w
- y-axis corresponds to v

- (b) has spurious local minima around (-1, -1.6)
- (c) has spurious local minima at bottom left

Success of SGD Assuming a Skip Connection to the Output

- Another important result - **Theorem 4** - of the paper is that, with modifying the residual architecture slightly by adding skip connection
- Showed that under light conditions, projected SGD with adequately many iterations will have a competitive objective value with a fixed linear predictor.

Success of SGD Assuming a Skip Connection to the Output

Theorem 4. $F(\mathbf{w}_t, \mathbf{v}_t, \theta_t) \leq \min_{\mathbf{u}:(\mathbf{u},\mathbf{0})\in\mathcal{M}_1} F_{\text{lin}}(\mathbf{u}) + \mathcal{O}\left(\frac{bl+r\sqrt{\log(1/\delta)}}{\sqrt{T}}\right)$

with the assumptions that

- $\mathcal{M} = \{(\mathbf{w}, \mathbf{v}, \theta) : (\mathbf{w}, \mathbf{v}) \in \mathcal{M}_1, \theta \in \mathcal{M}_2\}$ are closed convex sets
- θ and the loss function is l-Lipschitz in (w, v) over $\mathcal{M}_1$ and bounded in absolute value by r
- $\sqrt{\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2} \leq b$.

Summary

- Pros:
  - The paper is novel since there is no other work theoretically on residual networks with nonlinear activation units.
  - It is proven that for every local minimum of F satisfies upper-bounded by the global minimum of linear predictor.
  - Proving that training sufficiently with a projected SGD, residual networks with skip connection units are competitive with linear predictors.(**can be generalized into vector-valued outputs!**)
- Cons:
  - Upper-bound of F highly depends on the norm of V and w.
  - **Theorem 2 and 3** couldn't be proved for vector-valued outputs(In ML, problems are mostly have vector-valued outputs.)
  - In order to keep norms lower, you need to project gradients to a convex set.

End

Thank you for listening!

References

- [1]. Ohad Shamir. Are resnets provably better than linear predictors? arXiv preprint arXiv:1804.06739, 2018.
- [2]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385,2015.
- [3]. https://en.wikipedia.org/wiki/Lipschitz_continuity

Definition 1. ($\epsilon - SOPSP$). Let $\mathcal{M}$ be an open subset of the domain of $F(\mathbf{w}, V, \theta)$, on which $\nabla^2 F_\theta(\mathbf{w}, V)$ is μ_2-Lipschitz in $(\mathbf{w}, V)$. Then $(\mathbf{w}, V, \theta) \in \mathcal{M}$ is an ϵ-second-order partial stationary point ($\epsilon - \mathrm{SOPSP}$) of F on $\mathcal{M}$, if

$$\|\nabla F_\theta(\mathbf{w}, V)\| \leq \epsilon \text{ and } \lambda_{\min}\left(\nabla^2 F_\theta(\mathbf{w}, V)\right) \geq -\sqrt{\mu_2 \epsilon}$$

Lipschitz Continuity: Given two metric spaces (X, d_X) and (Y, d_Y), where d_X denotes the metric on the set X and d_Y is the metric on set Y, a function $f : X \to Y$ is called Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x_1 and x_2 in X

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2)$$

Intuitively, a Lipschitz continuous function is limited in how fast it can change: there exists a real number such that, for every pair of points on the graph of this function, the absolute value of the slope of the line connecting them is not greater than this real number; the smallest such bound is called the Lipschitz constant of the function.
(from Wikipedia)