Review of "Are ResNets Provably Better than Linear Predictors?"

Orhun Güley orhun.gueley@tum.de
03721364

1. Paper Summary

The paper analyzes the performance of residual networks with linear predictors. ResNet architecture consists of "residual units" in a mathematical form of $\mathbf{y} = \mathbf{x} + g_\Phi(\mathbf{x})$, which allows the network to focus on the "residual" of the previous layes's output, meaning that if there is nothing to learn, the identity function is used and more dept would not damage the model performance. As a result of this property, ResNets are empirically proved to be very successful while training extremely deep neural networks. The main focus of this paper is to rigorously prove that nonlinear deep ResNets don't have any local minimum which has a value above the value of the global minimum of a linear predictor, which allows network to select if it is necessary to learn the weights in that layer.

As preliminaries, a residual function and a linear predictor function is defined as follows;

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

Another important definition made in the paper is ϵ-SOPSP, which corresponds to the ϵ-second order stationary point. The term is defined as a point $(w, V, \theta) \in \mathcal{M}$ - $\mathcal{M}$ is a open subset in the domaion of $F(\mathbf{w}, V, \theta)$, on which $\nabla^2 F_\theta(\mathbf{w}, V)$ is μ_2-Lipschitz in $(\mathbf{w}, V)$ - where $\|\nabla F_\theta(\mathbf{w}, V)\| \leq \epsilon$ and $\lambda_{\min}\left(\nabla^2 F_\theta(\mathbf{w}, V)\right) \geq -\sqrt{\mu_2\epsilon}$. 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).

Theorem 3 and Corollary 1 are the main results of the paper, which are proven in 2 stages. In the first stage, they prove that for any point $(\mathbf{w}, V, \theta)$ such that $\mathbf{w} \neq 0$, in the cases which 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. In the second stage - Theorem 2 -, the case $w = 0$ is considered and objective value of F is upper-bounded with objective value of F_{lin} plus some value. In Theorem 3, those 2 results are combined and it is proved that under some considerable assumptions, that any ϵ-SOPSP of F is upper bounded by F_{lin} plus a polynomial expression multiplied by an ϵ-dependent function. This can be interpreted as in the case of $\epsilon = 0$, F will be only upper-bounded by only F_{lin}.

Paper also discusses the effects on norm and regularization on the performance of residual networks. It also draws attention that Theorem 3 depends on the norm of the point $(\|\mathbf{w}\|, \|V\|)$, since if the norm is sufficiently large, the polynomial term in the inequality will also be large. With an simple example, the paper shows that there will be a significant difference between objective values of residual network and a linear predictor. Paper also discusses adding a regularization term to the objective function, in order to prevent the norm increasing. Nonetheless, it is also shown graphically by examples that adding regularization might introduce new spurious local minima which doesn't exist in $F(w, V, \theta)$.

In the last part, paper shows that by modifying the residual architecture from $F(\mathbf{w}, V, \theta) = \mathbb{E}_{\mathbf{x},y}\left[\ell\left(\mathbf{w}^\top(\mathbf{x} + V f_\theta(\mathbf{x}))\,; y\right)\right]$ to $F(\mathbf{w}, \mathbf{v}, \theta) = \mathbb{E}_{\mathbf{x},y}\left[\ell\left(\mathbf{w}^\top\mathbf{x} + \mathbf{v}^\top f_\theta(\mathbf{x}); y\right)\right]$, running projected stochastic gradient descent with sufficiently many iterations results in a network competitive with linear predictor under mild conditions.

2. Proof Summary

The paper provides 4 proofs to show that ResNets are at least competitive as linear predictors. In **sub-section 1**, first three theorems are summarized since they are related. In **sub-section 2**, Theorem 4 is summarized.

2.1 Competitiveness with Linear Predictors

As stated in the summary section, the main result of this paper is Theorem 3 and Corollary 1, which is proved 2 two stages, using the findings of Theorem 1 and Theorem 2.

Proof of **Theorem 1** starts with **Lemma 1**, defining a matrix G in which $F(\mathbf{w}, V, \theta) - F_{lin}(\mathbf{w}^*)$ is upper-bounded by the inner product of vectorized G and gradient of F with respect to (w, V). By using **Lemma 1** and **Cauchy-Schwartz** inequality, they get the following inequality;
$\|G\|_{Fr} \cdot \|\nabla F_\theta(\mathbf{w}, V)\| \geq \|G\| \cdot \|\nabla F_\theta(\mathbf{w}, V)\| \geq F(\mathbf{w}, V, \theta) - F_{lin}(\mathbf{w}^*)$

By dividing both sides by Frobenius norm of matrix G and upper-bounding the denominator, they show that **Theorem 1** holds for any point (w, V, θ) that $w \neq 0$ as follows;

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

In **Theorem 2**, paper analyzes the case where $w = 0$. To find an upper-bound for $F(\mathbf{0}, V, \theta) - F_{lin}(\mathbf{w}^*)$, paper combines two inequalities. To get to the first equation, paper proposes **Lemma 2**, where it defines an symmetric real-valued square matrix M of form $M = \begin{pmatrix} b & \mathbf{r}^\top \\ \mathbf{r} & \mathbf{0} \end{pmatrix}$ where b is some scalar, r is a vector, and all entries of M other than the first row and column are 0.

Then the minimal eigenvalue $\lambda_{\min}$ of M is non-positive, and satisfies $\|\mathbf{r}\|^2 = |b\lambda_{\min}| + \lambda_{\min}^2$. Afterwards, paper shows that M is a submatrix of $\nabla^2 F_\theta(\mathbf{0}, V)$, that's why it should follow the inequality $\lambda_{\min}\left(\nabla^2 F_\theta(\mathbf{0}, V)\right) \leq \lambda_{\min}(M) \leq 0$. By using that equation and

Lemma 2, we get the first inequality as follows;

$$\|\mathbf{r}\|^2 \leq \left|\lambda_{\min}\left(\nabla^2 F_\theta(\mathbf{0},V)\right)\right| \cdot \left\|\frac{\partial^2}{\partial \mathbf{w}^2} F_\theta(\mathbf{0},V)\right\| + \lambda_{\min}\left(\nabla^2 F_\theta(\mathbf{0},V)\right)^2 \text{ where } \mathbf{r} = \mathbb{E}_{\mathbf{x},y}\left[\ell'(0;y) f_\theta(\mathbf{x})\right]$$

By triangle and Cauchy-Shwartz inequalities, paper gets to the second equation $\|\nabla F_\theta(\mathbf{0},V)\| + \|V\| \cdot \|\mathbf{r}\| \geq \|\nabla F_{lin}(\mathbf{0})\|$.By combining those two equations, the **Theorem 2** holds as follows;

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

In **Theorem 3**, paper combines 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.
- For 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 $\epsilon \geq 0$

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

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

As it is stated in the paper summary and also the original paper, 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

2.2 Success of SGD Assuming a Skip Connection to the Output

The other important result - **Theorem 4** - of the paper is that, with modifying the residual architecture slightly by adding skip connection, they show that under light conditions, projected SGD with adequately many iterations is will have a competitive objective value with a fixed linear predictor. The key observation of the paper for the projected SGD algorithm is that the updated weights of the current state in algorithm is identical to the next iterations input weights. With this observations, paper uses **Theorem 5** to find an upper bound for objective value of F, which is as follows;
$F\left(\mathbf{w}_t, \mathbf{v}_t, \theta_t\right) \leq \min_{\mathbf{u}:(\mathbf{u},\mathbf{0}) \in \mathcal{M}_1} F_{\text{lin}}(\mathbf{u}) + \mathcal{O}\left(\frac{bl + r\sqrt{\log(1/\delta)}}{\sqrt{T}}\right)$
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, and $\sqrt{\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2} \leq b$.

3. Review Summary

Paper is a good contribution to the field of theoretical machine learning. The paper analyzes the performance of ResNets with non-linearity, compared to linear predictors. It is men-

tioned that the existing work related to rigorous theoretical analysis on residual networks all refer to linear architectures, which makes the work done in this paper original.

3.1 Pros

1) Novel since there is no other theoretical work on nonlinear ResNets.
2) Proved any local minimum of F upper-bounded by $F_{lin}(w^*)$.
3) Proved ResNets with skip connection units are competitive with linear predictors.(**can be generalized into vector-valued outputs!**)

3.2 Cons

1) Upper-bound of F highly depends on the norm of V and w.
2) **Th. 2&3** couldn't be proved for vector-valued outputs(In ML, problems are mostly have vector-valued outputs.)
3) In order to keep norms lower, you need to project gradients to a convex set.

4. Detailed Evaluation

The novelty and significance of the paper can be examined as follows. In a problem setting point of view, the paper is novel because there haven't been many existing research done o rigorous theoretical understanding of training nonlinear residual networks. This is an important property which makes the work done in this paper original and significant because nonlinear activation functions introduce non-convexity to deep neural networks. The paper clearly advances our theoretical understanding of residual networks.

The paper is clearly written and well-organized. All the steps in the paper is written in detail. Proofs of the theorems are also written in detail and easy to follow. The given examples in **Section 4** can be easily understood by a well-informed person in the machine learning field.

The paper is technical and all the theorems and corollaries in the paper is supported by high level ideas. I believe the paper is nearly complete piece of work. The only downside is **Th.2** and **Th.3** is not proved for vector-valued outputs. But after adding skip connections as it is in **Sec.5**, they are able to generalize the **Th.4** for vector-valued outputs as well, which I believe is a very important result. The authors of the paper are very clear about the strong and weak spots of the paper. They approach their theorems in a critical way and they fairly mention why in some cases the theorems won't hold. In **Remark 2**, they discuss the generalization of their theorems to vector-valued outputs and they indicate that it is not clear for them to prove **Th.2** and **Th.3** - a variant of **Th.1** is proved in Appendix - by taking vector-valued arguments. They even give examples and also graphically illustrate their examples to enhance the intuition. I find the **Ex.1** and **Remark 3** very useful. In **Ex. 1**, they show a counter example for the idea that $\|\mathbf{w}\|, \|V\|$ is an artifact of the analysis, and any ϵ-SOPSP will be competitive with a linear predictor. In the **Ex.1**, even though the Hessian is very close to zero, the difference between linear predictor and residual network was significant. Afterwards, they discuss to idea of adding a regularization term to

the objective function in order to prevent the norms of $\|\mathbf{w}\|$ and $\|V\|$ getting larger values. They show in a graphical figure that this will introduce new spurious local minimums to the objective.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385,2015.