# ResNEsts and DenseNEsts: Block-based DNN Models with Improved Representation Guarantees

**Kuan-Lin Chen**[1], **Ching-Hua Lee**[1], **Harinath Garudadri**[2], **and Bhaskar D. Rao**[1]
[1]Department of Electrical and Computer Engineering, [2]Qualcomm Institute
University of California, San Diego

**Contact Information:**
University of California, San Diego
9500 Gilman Drive #0436, La Jolla, CA 92093

Email: kuc029@ucsd.edu

UC San Diego JACOBS SCHOOL OF ENGINEERING Electrical and Computer Engineering

QUALCOMM INSTITUTE

NEURAL INFORMATION PROCESSING SYSTEMS

## Abstract

- We propose ResNEsts, i.e., Residual Nonlinear Estimators, by simply dropping nonlinearities at the last residual representation from standard ResNets.
- Wide ResNEsts with bottleneck blocks can always guarantee a very desirable training property, i.e., adding more blocks does not decrease performance.
- We propose DenseNEsts, i.e., Densely connected Nonlinear Estimators and show that their theoretical guarantees are superior to ones obtained in ResNEsts.

## 1 ResNEsts and augmented ResNEsts

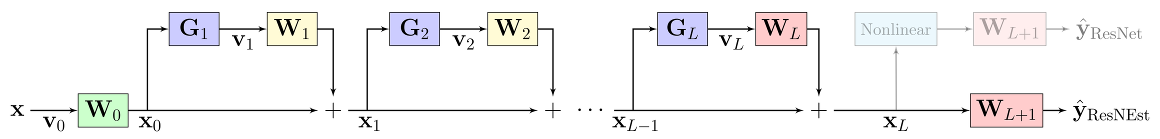

**Figure 1:** A generic vector-valued ResNEst that has a chain of $L$ residual blocks (or units). Different from the ResNet architecture using pre-activation residual blocks in the literature [1], our ResNEst architecture drops nonlinearities at $\mathbf{x}_L$ so as to reveal a linear relationship between the output $\hat{\mathbf{y}}_{\text{ResNEst}}$ and the features $\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_L$.

### 1.1 Dropping nonlinearities and expanding the input space

The proposed ResNEst model employs the following input-output relationship for the $i$-th residual block in Figure 1:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \mathbf{W}_i \mathbf{G}_i\left(\mathbf{x}_{i-1}; \boldsymbol{\theta}_i\right). \quad (1)$$

The term $\mathbf{W}_i\mathbf{G}_i$ is a composition of a nonlinear function $\mathbf{G}_i$ and a linear transformation, which is generally known as a residual function. $\mathbf{W}_i \in \mathbb{R}^{M \times K_i}$ forms a linear transformation and we consider $\mathbf{G}_i\left(\mathbf{x}_{i-1}; \boldsymbol{\theta}_i\right) : \mathbb{R}^M \mapsto \mathbb{R}^{K_i}$ as a function implemented by a neural network with parameters $\boldsymbol{\theta}_i$ for all $i \in \{1, 2, \cdots, L\}$. We define the expansion $\mathbf{x}_0 = \mathbf{W}_0\mathbf{x}$ for the input $\mathbf{x} \in \mathbb{R}^{N_{in}}$ to the ResNEst using a linear transformation with a weight matrix $\mathbf{W}_0 \in \mathbb{R}^{M \times K_0}$. The output $\hat{\mathbf{y}}_{\text{ResNEst}} \in \mathbb{R}^{N_o}$ (or $\hat{\mathbf{y}}_{L\text{-ResNEst}}$ to indicate $L$ blocks) of the ResNEst is defined as $\hat{\mathbf{y}}_{L\text{-ResNEst}}(\mathbf{x}) = \mathbf{W}_{L+1}\mathbf{x}_L$ where $\mathbf{W}_{L+1} \in \mathbb{R}^{N_o \times M}$.

- $M$ is the expansion factor.
- $N_o$ is the output dimension of the network.

### 1.2 Basis function modeling and the coupling problem

Because the ResNEst now reveals a linear relationship between the output and the features, we have:

$$\hat{\mathbf{y}}_{L\text{-ResNEst}}(\mathbf{x}) = \mathbf{W}_{L+1} \sum_{i=0}^{L} \mathbf{W}_i \mathbf{v}_i(\mathbf{x}) \quad (2)$$

where

$$\mathbf{v}_i(\mathbf{x}) = \mathbf{G}_i\left(\mathbf{x}_{i-1}; \boldsymbol{\theta}_i\right) = \mathbf{G}_i\left(\sum_{j=0}^{i-1} \mathbf{W}_j \mathbf{v}_j; \boldsymbol{\theta}_i\right). \quad (3)$$

We propose to utilize the basis function modeling point of view in the ResNEst and analyze the following ERM problem:

$$(\mathrm{P}_\phi) \min_{\mathbf{W}_L, \mathbf{W}_{L+1}} \mathcal{R}\left(\mathbf{W}_L, \mathbf{W}_{L+1}; \phi\right) \quad (4)$$

where

$$\mathcal{R}\left(\mathbf{W}_L, \mathbf{W}_{L+1}; \phi\right) = \frac{1}{N} \sum_{n=1}^{N} \ell\left(\hat{\mathbf{y}}^{\phi}_{L\text{-ResNEst}}\left(\mathbf{x}^n\right), \mathbf{y}^n\right) \quad (5)$$

for any fixed feature finding weights $\phi$.

**Remark 1.** *Since the set of all local minima of $(P_\phi)$ using any possible features is a superset of the set of all local minima of the original ERM problem (P), any characterization of $(P_\phi)$ can then be translated to (P).*

**Assumption 1.** $\sum_{n=1}^{N} \mathbf{v}_L\left(\mathbf{x}^n\right) \mathbf{y}^{nT} \neq \mathbf{0}$ *and* $\sum_{n=1}^{N} \mathbf{v}_L\left(\mathbf{x}^n\right) \mathbf{v}_L\left(\mathbf{x}^n\right)^T$ *is full rank.*

**Proposition 1.** *If $\ell$ is the squared loss and Assumption 1 is satisfied, then*

*(a) the objective function of $(P_\phi)$ is non-convex and non-concave;*

*(b) every critical point that is not a local minimizer is a saddle point in $(P_\phi)$.*

### 1.3 Bounding empirical risks via augmentation

To avoid the coupling problem in ResNEsts, an $L$-block A-ResNEst introduces another set of parameters $\{\mathbf{H}_i\}_{i=0}^{L}$ to replace every bilinear map on each feature in (2) with a linear map:

$$\hat{\mathbf{y}}_{L\text{-A-ResNEst}}(\mathbf{x}) = \sum_{i=0}^{L} \mathbf{H}_i \mathbf{v}_i(\mathbf{x}). \quad (6)$$

**Assumption 2.** *The loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$ is differentiable and convex in $\hat{\mathbf{y}}$ for any $\mathbf{y}$.*

**Proposition 2.** *Let $\left(\mathbf{H}_0^*, \cdots, \mathbf{H}_L^*\right)$ be any local minimizer of the following optimization problem:*

$$(PA_\phi) \min_{\mathbf{H}_0, \cdots, \mathbf{H}_L} \mathcal{A}\left(\mathbf{H}_0, \cdots, \mathbf{H}_L; \phi\right) \quad (7)$$

*where $\mathcal{A}\left(\mathbf{H}_0, \cdots, \mathbf{H}_L; \phi\right) = \frac{1}{N}\sum_{n=1}^{N} \ell\left(\hat{\mathbf{y}}^{\phi}_{L\text{-A-ResNEst}}\left(\mathbf{x}^n\right), \mathbf{y}^n\right)$. If Assumption 2 is satisfied, then the optimization problem in (7) is convex and*

$$\epsilon = \mathcal{R}\left(\mathbf{W}_L^*, \mathbf{W}_{L+1}^*; \phi\right) - \mathcal{A}\left(\mathbf{H}_0^*, \cdots, \mathbf{H}_L^*; \phi\right) \geq 0 \quad (8)$$

*for any local minimizer $\left(\mathbf{W}_L^*, \mathbf{W}_{L+1}^*\right)$ of $(P_\phi)$ using arbitrary feature finding parameters $\phi$.*

### 1.4 Condition for strictly improved representations

**Question 1.** *What properties are fundamentally required for features to strictly improve the representation over blocks?*

A fundamental answer is they need to be at least *linearly unpredictable*. Note that $\mathbf{v}_i$ must be linearly unpredictable by $\mathbf{v}_0, \cdots, \mathbf{v}_{i-1}$ if $\mathcal{A}\left(\mathbf{H}_0^*, \mathbf{H}_1^*, \cdots, \mathbf{H}_{i-1}^*, \mathbf{0}, \cdots, \mathbf{0}, \phi^*\right) > \mathcal{A}\left(\mathbf{H}_0^*, \mathbf{H}_1^*, \cdots, \mathbf{H}_i^*, \mathbf{0}, \cdots, \mathbf{0}, \phi^*\right)$ for any local minimum $\left(\mathbf{H}_0^*, \cdots, \mathbf{H}_L^*, \phi^*\right)$ in (PA). The residual representation $\mathbf{x}_i$ is not strictly improved from the previous representation $\mathbf{x}_{i-1}$ if the feature $\mathbf{v}_i$ is linearly predictable by the previous features.

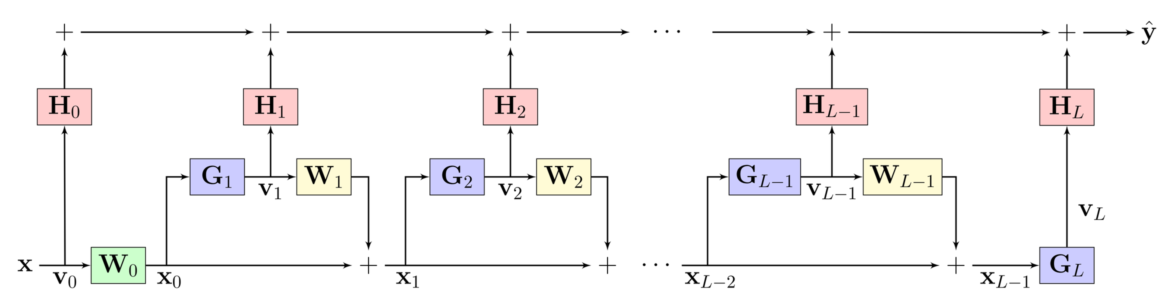

**Figure 2:** The proposed augmented ResNEst or A-ResNEst.

## 2 Wide ResNEsts with bottleneck residual blocks

**Assumption 3.** $M \geq N_o$.

**Assumption 4.** *The linear inverse problem $\mathbf{x}_{L-1} = \sum_{i=0}^{L-1} \mathbf{W}_i \mathbf{v}_i$ has a unique solution.*

**Theorem 1.** *If Assumption 2 and 3 are satisfied, then the following two properties are true in $(P_\phi)$ under any $\phi$ such that Assumption 4 holds:*

*(a) every critical point with full rank $\mathbf{W}_{L+1}$ is a global minimizer;*

*(b) $\epsilon = 0$ for every local minimizer.*

**Remark 2.** *Let Assumption 2 and 3 be true. Any local minimizer of (P) such that Assumption 4 is satisfied guarantees*

*(a) monotonically improved (no worse) residual representations over blocks;*

*(b) every residual representation is better than the input representation in the linear prediction sense.*

**Corollary 1.** *Let Assumption 2 and 3 be true. Any local minimum of $(P_\alpha)$ is smaller than or equal to any local minimum of $(P_\beta)$ under Assumption 4 for any $\boldsymbol{\alpha} = \{\mathbf{W}_{i-1}, \boldsymbol{\theta}_i\}_{i=1}^{L_\alpha}$ and $\boldsymbol{\beta} = \{\mathbf{W}_{i-1}, \boldsymbol{\theta}_i\}_{i=1}^{L_\beta}$ where $L_\alpha$ and $L_\beta$ are positive integers such that $L_\alpha > L_\beta$.*

**Corollary 2.** *Let $\left(\mathbf{W}_0^*, \cdots, \mathbf{W}_{L+1}^*, \boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_L^*\right)$ be any local minimizer of (P) and $\phi^* = \{\mathbf{W}_{i-1}^*, \boldsymbol{\theta}_i^*\}_{i=1}^{L}$. If Assumption 2, 3 and 4 are satisfied, then*

*(a) $\mathcal{R}\left(\mathbf{W}_0^*, \cdots, \boldsymbol{\theta}_L^*\right) \leq \min_{\mathbf{A} \in \mathbb{R}^{N_o \times N_{in}}} \frac{1}{N}\sum_{n=1}^{N} \ell\left(\mathbf{A}\mathbf{x}^n, \mathbf{y}^n\right)$;*

*(b) the above inequality is strict if $\mathcal{A}\left(\mathbf{H}_0^*, \mathbf{0}, \cdots, \mathbf{0}, \phi^*\right) > \mathcal{A}\left(\mathbf{H}_0^*, \cdots, \mathbf{H}_L^*, \phi^*\right)$.*

**Theorem 2.** *If $\ell$ is the squared loss, and Assumption 1 and 3 are satisfied, then the following two properties are true at every saddle point of $(P_\phi)$ under any $\phi$ such that Assumption 4 holds:*

*(a) $\mathbf{W}_{L+1}$ is rank-deficient;*

*(b) there exists at least one direction with strictly negative curvature.*

## 3 DenseNEsts are wide ResNEsts with bottleneck residual blocks equipped with orthogonalities

For an $L$-block DenseNEst, we define the $i$-th dense block as a function $\mathbb{R}^{M_{i-1}} \mapsto \mathbb{R}^{M_i}$ of the form

$$\mathbf{x}_i = \mathbf{x}_{i-1} \copyright \mathbf{Q}_i\left(\mathbf{x}_{i-1}; \boldsymbol{\theta}_i\right) \quad (9)$$

for $i = 1, 2, \cdots, L$ where the dense function $\mathbf{Q}_i$ is a general nonlinear function; and $\mathbf{x}_i$ is the output of the $i$-th dense block. For all $i \in \{1, 2, \cdots, L\}$, $\mathbf{Q}_i(\mathbf{x}_{i-1}; \boldsymbol{\theta}_i) : \mathbb{R}^{M_{i-1}} \mapsto \mathbb{R}^{D_i}$ is a function implemented by a neural network with parameters $\boldsymbol{\theta}_i$ where $D_i = M_i - M_{i-1} \geq 1$ with $M_0 = N_{in} = D_0$. The output of a DenseNEst is defined as $\hat{\mathbf{y}}_{\text{DenseNEst}} = \mathbf{W}_{L+1}\mathbf{x}_L$ for $\mathbf{W}_{L+1} \in \mathbb{R}^{N_o \times M_L}$.

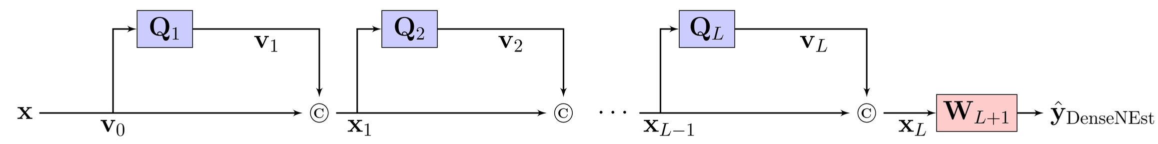

**Figure 3:** A generic vector-valued DenseNEst that has a chain of $L$ dense blocks (or units). The symbol "©" represents the concatenation operation.

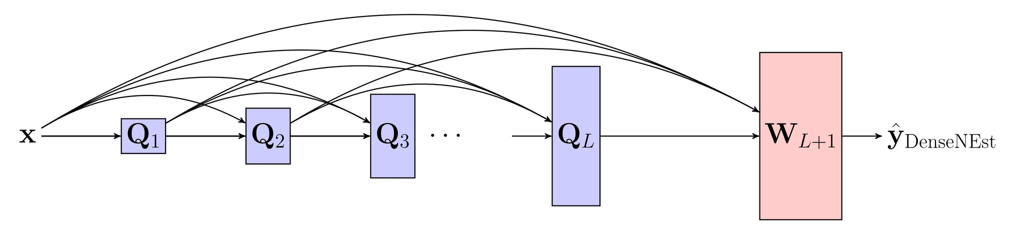

**Figure 4:** An equivalence to Figure 3 emphasizing the growth of the input dimension at each block.

The ERM problem (PD) for the DenseNEst is defined on $\left(\mathbf{W}_{L+1}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_L\right)$. The DenseNEst ERM problem for any fixed features, denoted as $(\mathrm{PD}_\phi)$, is given by

$$(\mathrm{PD}_\phi) \min_{\mathbf{W}_{L+1}} \mathcal{D}\left(\mathbf{W}_{L+1}; \phi\right) \quad (10)$$

where $\mathcal{D}\left(\mathbf{W}_{L+1}; \phi\right) = \frac{1}{N}\sum_{n=1}^{N} \ell\left(\hat{\mathbf{y}}^{\phi}_{L\text{-DenseNEst}}\left(\mathbf{x}^n\right), \mathbf{y}^n\right)$.

**Proposition 3.** *If Assumption 2 is satisfied, then any local minimum of (PD) is smaller than or equal to the minimum empirical risk given by any linear predictor of the input.*

**Proposition 4.** *Given any DenseNEst $\hat{\mathbf{y}}_{L\text{-DenseNEst}}$, there exists a wide ResNEst with bottleneck residual blocks $\hat{\mathbf{y}}^{\phi}_{L\text{-ResNEst}}$ such that $\hat{\mathbf{y}}^{\phi}_{L\text{-ResNEst}}(\mathbf{x}) = \hat{\mathbf{y}}_{L\text{-DenseNEst}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^{N_{in}}$. If, in addition, Assumption 2 and 3 are satisfied, then $\epsilon = 0$ for every local minimizer of $(P_\phi)$.*

## Acknowledgements

This work was supported in part by NSF under Grant CCF-2124929 and Grant IIS-1838830, in part by NIH/NIDCD under Grant R01DC015436, Grant R21DC015046, and Grant R33DC015046, in part by Halıcıoğlu Data Science Institute, and in part by Wrethinking, the Foundation.

## References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.