Understanding Deep Neural Networks through the lens of their Non-Linearity

Quentin Bouniot[1] Ievgen Redko[2] Anton Mallasto[3] Charlotte Laclau[1]
Karol Arndt[4] Oliver Struckmeier[4] Markus Heinonen[4] Ville Kyrki[4]
Samuel Kaski[4,5]

[1]Telecom Paris [2]Noah's Ark Lab [3]Smartly.io [4]Aalto University [5]University of Manchester

Motivations

Non-linearity is at the heart of DNNs

- **Universal function approximators** thanks to non-linearity.
- Mainly introduced through **activation functions**.

No such notion of quantifying non-linearity exists in the literature.

- Research mainly focus on quantifying **expressive power** of DNNs.

Goal: Measure non-linearity of activation functions **from data distribution**

Outline

1. Quantifying Non-linearity
2. Journey through DNNs History
3. Additional Results

Outline

General idea

Measure *non-linearity* as *lack of linearity* through *Optimal Transport* (OT)

- OT is a theoretically sounded tool when **comparing distributions**.
- We want a **bounded score** to be able to compare behaviors.

General idea

Measure *non-linearity* as *lack of linearity* through *Optimal Transport* (OT)

- We know the **closed-form solution** of the OT problem for random variables (RVs) following **normal distributions**.
- For any RVs X and Y, if $Y = TX$ with T Positive Semi-Definite (PSD) matrix, then **the solution of OT problem is exactly the one of their normal approximations** ($N_X \sim \mathcal{N}(\mu(X), \Sigma(X))$ and $N_Y \sim \mathcal{N}(\mu(Y), \Sigma(Y))$).
- We obtain an **upper bound** on the difference of the two OT problems.
- We can define the **affinity score** using this bound.

Identifiability

Theorem (Smith & Knott[1])

Let $X \sim \mu$ and $T(x) = \nabla\phi(x)$ for a convex function ϕ. Then T is the unique OT map between μ and $T_{\#}\mu$.

Corollary

If $Y = TX$ where T is a PSD affine transformation, then T is the unique OT map from X to Y.

- Uniqueness OT maps expressed as **gradients of convex functions**.
- Whenever two RVs are linked through an **affine transformation**, then the solution to the OT problem between them is **exactly the affine transformation**.

[1] Cyril S Smith and Martin Knott. "Note on the optimal transportation of distributions". In: *Journal of Optimization Theory and Applications* 52.2 (1987), pp. 323–329.

Characterization

Theorem

For any RVs X, Y such that $Y = TX$ with T a PSD matrix. Let N_X and N_Y be their normal approximations. Then $W_2(N_X, N_Y) = W_2(X, Y)$ and $T = T_{\text{aff}}$ where T_{aff} is the OT map between N_X and N_Y and can be calculated in closed-form:

$$T_{\text{aff}}(x) = Ax + b, \quad A = \Sigma(Y)^{\frac{1}{2}} \left(\Sigma(Y)^{\frac{1}{2}} \Sigma(X) \Sigma(Y)^{\frac{1}{2}}\right)^{-\frac{1}{2}} \Sigma(Y)^{\frac{1}{2}}, \quad b = \mu(Y) - A\mu(X). \tag{1}$$

- When X and Y differ by an **affine transformation**, the OT solution can be computed in closed-form using their normal approximations, **no matter how complicated X and Y are**.

Upper bound

Theorem (Gelbrich bound[2])

Let X, Y be RVs and N_X, N_Y be their normal approximations. Then, $W_2(N_X, N_Y) \leq W_2(X, Y)$.

Proposition

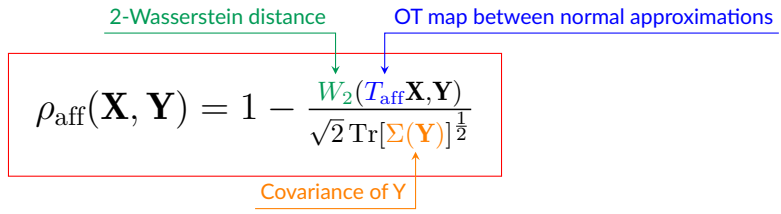
Let X, Y be RVs and N_X, N_Y be their normal approximations. For T_{aff} as in (1),

$$W_2(T_{\text{aff}}X, Y) \leq \sqrt{2}\,\mathrm{Tr}\left[\Sigma(Y)\right]^{\frac{1}{2}}. \tag{2}$$

- The difference in the cost between the two OT problems **increases when the map becomes non-linear**.
- We have equality iff RVs are linked through an **affine transformation**.

[2] Matthias Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Mathematische Nachrichten* 147.1 (1990), pp. 185–203.

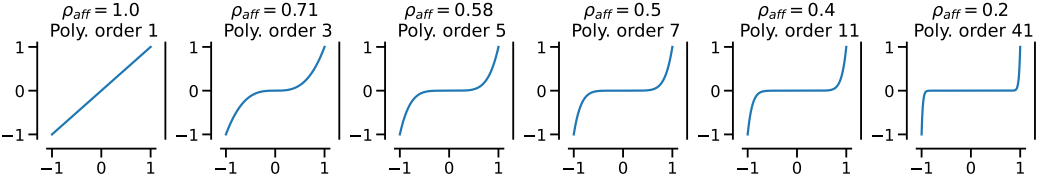
Affinity Score



- ρ_{aff} describes how much Y differs from being a *PSD affine transformation of* X.
- $0 \leq \rho_{\text{aff}}(X, Y) \leq 1$, and $\rho_{\text{aff}}(X, Y) = 1 \Leftrightarrow Y = T_{\text{aff}}X$.

Toy Examples

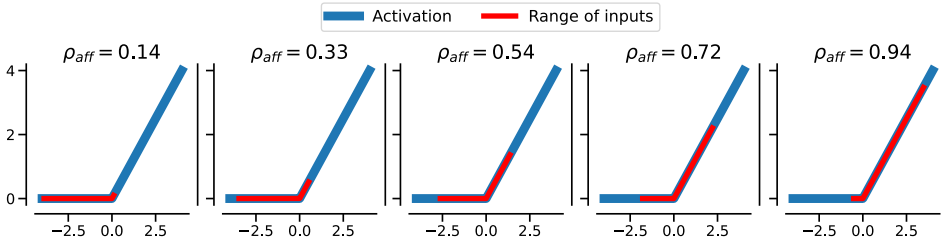
Affinity scores for simple polynomial functions



- Affinity score decreases as the transformation becomes **more non-linear**.

ReLU example

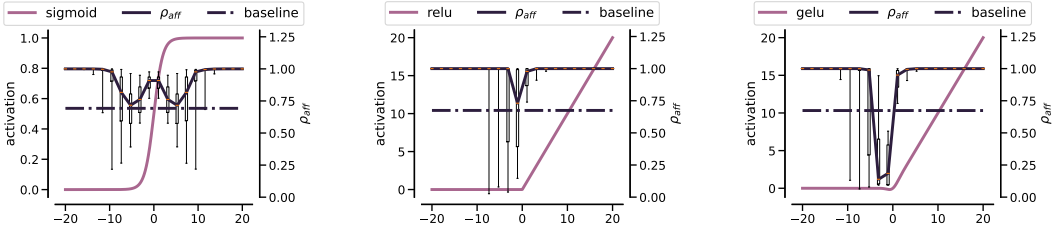
Affinity scores throughout the input domain of ReLU



- Affinity scores will *vary* depending on the **input domain considered**.
- For ReLU, **high** ρ_{aff} **values** in the linear part of the transformation.

Examples of Activation Functions

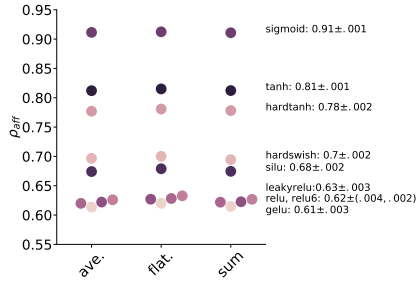
Affinity scores over input domain of activation functions



- $\mathbf{X} \sim \mathcal{N}(\mu, \sigma)$, with μ sliding over the domain and multiple σ for each μ.
- $\rho_{\text{aff}}(\mathbf{X}, f(\mathbf{X}))$ for popular activation functions f.
- Activation functions can be characterized by **the lowest score achieved** and **the range of non-linearity**.

Robustness to Pooling

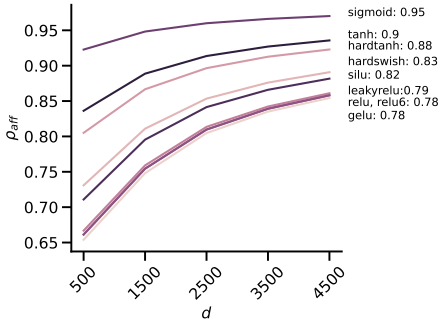
Affinity scores are robust to pooling

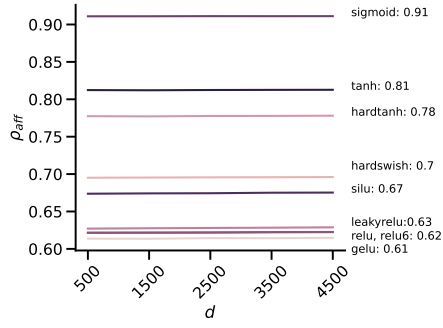


- ▶ Manipulating 4-order tensor is computationally expensive
- ✓ Averaging over a dimension preserve affinity scores

Covariance estimation

Shrinkage of the covariance makes it robust to sample size





✓ **Ledoit-Wolfe shrinkage**[3] of the covariance gives stable results for affinity scores.

[3] Olivier Ledoit and Michael Wolf. **"Honey, I shrunk the sample covariance matrix"**. In: *Journal of Portfolio Management* 30.4 (2004), pp. 110–119.

Outline

Non-linearity signature

Notations

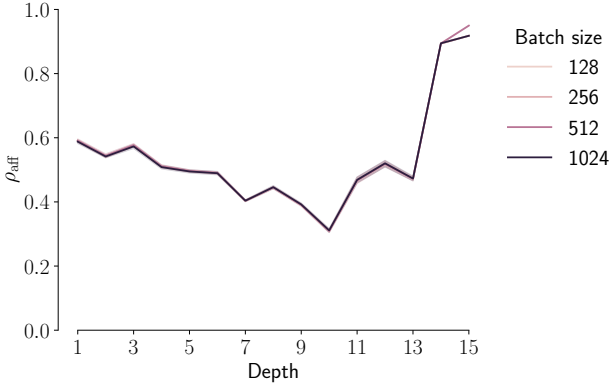
- Define a neural network N as a *composition of layers* F_i: $N = F_L \odot ... \odot F_i \ ... \odot F_1 = \bigodot_{k=1,...,L} F_k$ where $\odot$ stands for a composition.
- Each layer F_i is a function $F_i : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times c}$ whose outputs $F_i(\mathbf{X}_i)$ are inputs of the following layer F_{i+1}. Usual F_i include convolution, feedforward, pooling or activation functions.
- Define a *finite set of common activation functions* $\mathcal{A} := \{\sigma | \sigma : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times c}\}$
- Let r be a *pooling operation* such that $r : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{c}$

Non-linearity signature of N given $\mathbf{X}$:

$$\rho_{\text{aff}}(N; \mathbf{X}) = \{\rho_{\text{aff}}(r(\mathbf{X}_i), \sigma(r(\mathbf{X}_i))), \forall \sigma \in F_i \cap \mathcal{A}, i \in \{1, \ldots, L\}\}$$

Empirical Properties

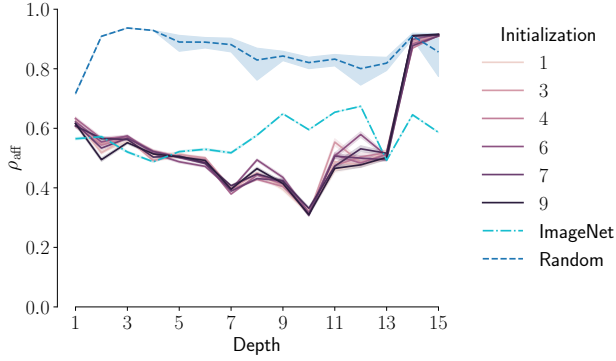
Stability with respect to the choice of batch size



- VGG16 on CIFAR10, ρ_{aff} computed with *different batch size*
- ✓ Stable with respect to **batch size**

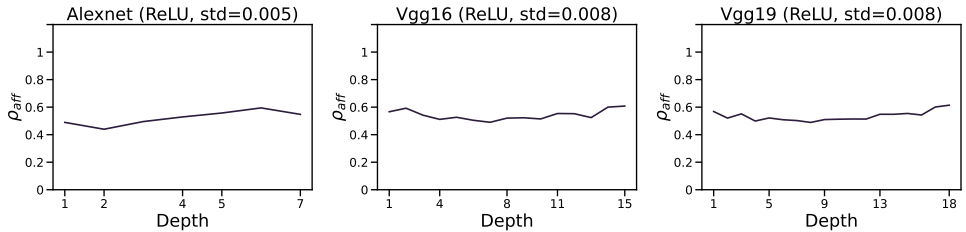
Empirical Properties

Impact of DNN's weights



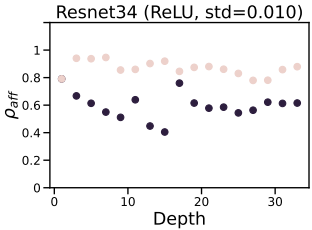
- ▶ VGG16 with *different initialization* on CIFAR10
- ✓ Stable with respect to same training but **different random seeds**
- ✓ Difference between **learned weights**

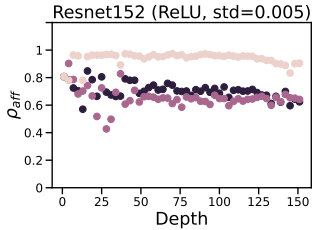
Early Convnets

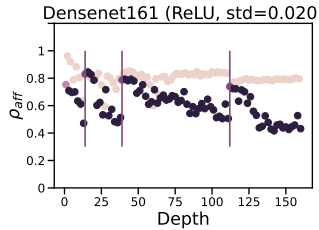


- Early convnets had **tiny variations** in non-linearity propagation.

Deeper Networks

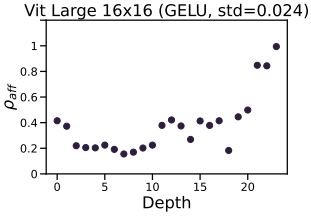


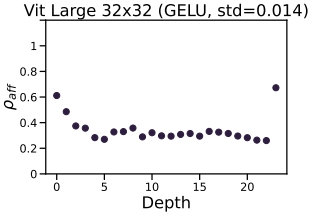


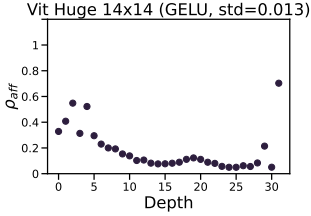


- Different color codes stand for *distinct* activation functions appearing *repeatedly* in each block (*e.g.* every first ReLU in residual blocks for ResNet).
- Deeper networks with *residual connections* have a **shaking effect** in their non-linearity signatures.

Vision Transformers

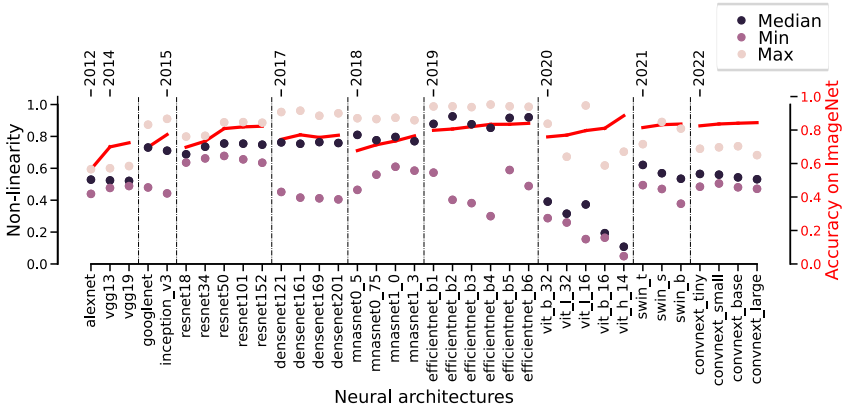






- Activation functions only present in their MLP blocks.
- **Highly non-linear** compared to convnets.

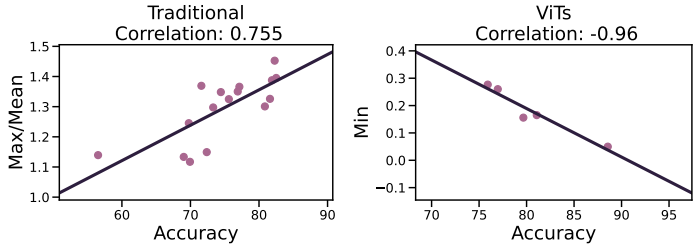
Throughout DNNs Architectures



- **Affinity scores statistics** and **Accuracy** (in red) throughout DNNs architectures.
- Before ViTs: **max and median** values are increasing, also **gap between min and max**.
- Within ViTs: Trend of **decreasing min values**

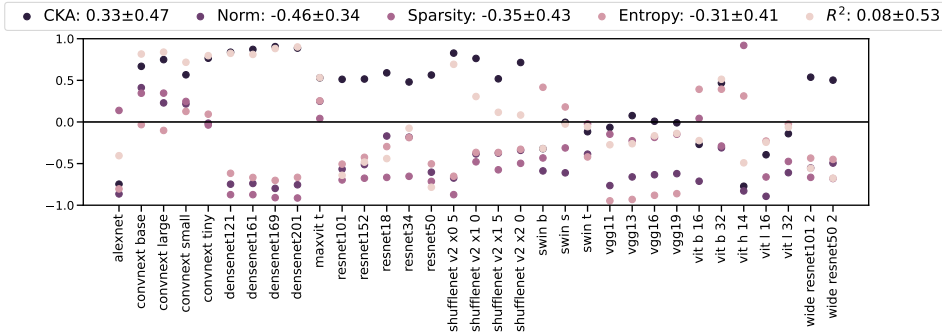
Outline

Correlation with Accuracy



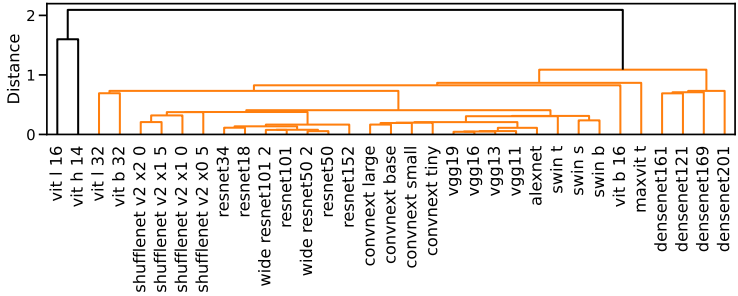
- We separate architectures into semantically meaningful groups: **Traditional architectures** (Alexnet, VGGs, ResNets and DenseNets) and **ViTs**.
- Confirms **shaking effect** for traditional models.
- Clear trend toward **more non-linearity in ViTs**.

Unique Measure



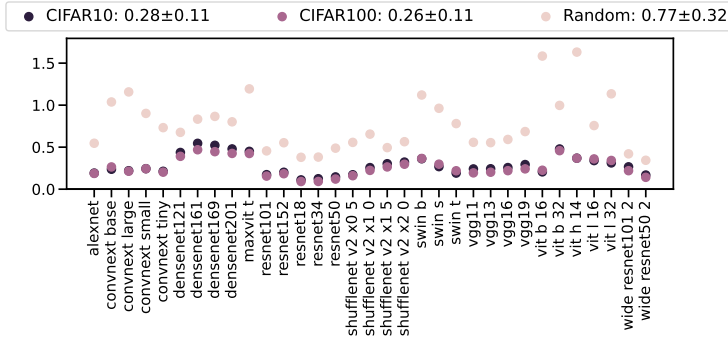
- **No other criterion consistently correlates with the affinity score** across 33 architectures used in our test.

Clustering of architectures



- Clustering of the architectures using **the pairwise Dynamic Time Window (DTW) distances** between non-linearity signatures.

Deviations between datasets



- **Deviations to ImageNet of different datasets** (CIFAR10, CIFAR100, random data), for each architecture.

Take-Home Message

Understanding Deep Neural Networks Through the Lens of their Non-Linearity[4]

- ✓ First theoretical sound tool to measure non-linearity in DNNs
- ✓ Different developments in Deep Learning can be understood through the prism of non-linearity
- ✓ Variety of potential applications

[4] Quentin Bouniot et al. "Understanding deep neural networks through the lens of their non-linearity". In: *arXiv preprint arXiv:2310.11439* (2023).

Thank you for listening !

Do not hesitate to contact us if you have questions.

 Quentin Bouniot et al. “Understanding deep neural networks through the lens of their non-linearity”. In: *arXiv preprint arXiv:2310.11439* (2023).