From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Networks with Affine Optimal Transport

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** which are the common ingredients between architectures.

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**

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.

Affinity Score

2-Wasserstein distance

OT map between normal approximations

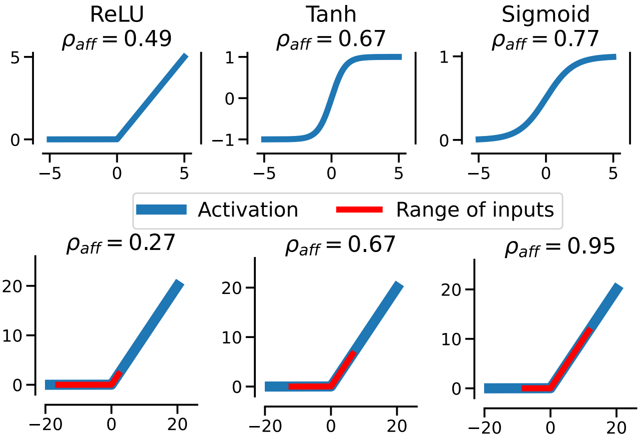
$$\rho_{\text{aff}}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{W_2(T_{\text{aff}}\mathbf{X}, \mathbf{Y})}{\sqrt{2\,\mathrm{Tr}[\Sigma(\mathbf{Y})]}}$$

Covariance of Y

- 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 $\rho_{\text{aff}}(X, Y) = 1 \Leftrightarrow Y = T_{\text{aff}}X$.

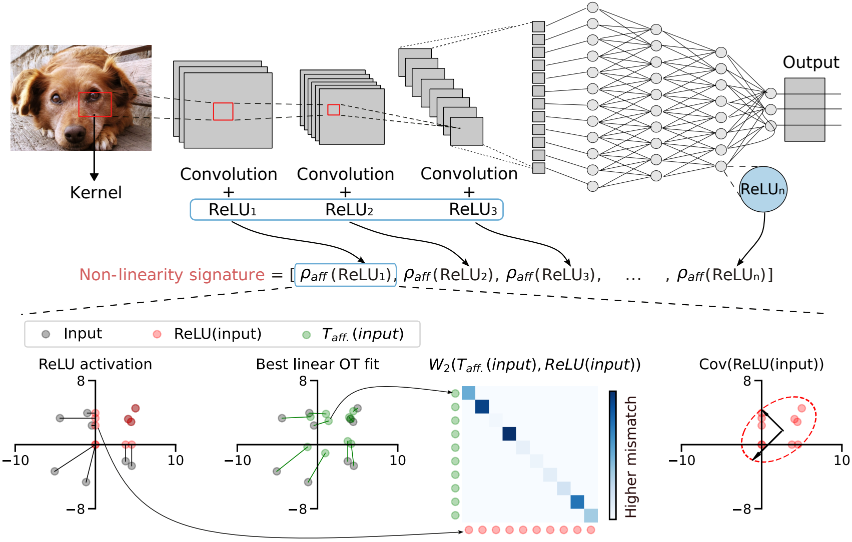
ReLU example

Affinity scores throughout the input domain of ReLU

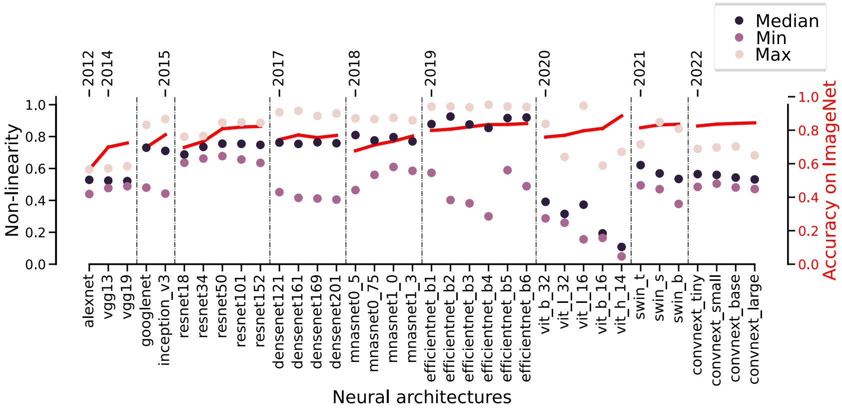


- Affinity scores will *vary* depending on the **input domain considered**.
- For ReLU, **high** ρ_{aff} **values** in the linear part of the transformation.

Non-linearity signature



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**

Take-Home Message

From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Networks with Affine Optimal Transport[1]

- ✓ 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

[1] Quentin Bouniot et al. "From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Networks with Affine Optimal Transport". In: *arXiv preprint arXiv:2310.11439* (2023).

Thank you for listening !

Do not hesitate to contact us if you have questions.

[1] Quentin Bouniot et al. “From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Networks with Affine Optimal Transport”. In: *arXiv preprint arXiv:2310.11439* (2023).