Invariance and Stability of Deep Convolutional Representations

Alberto Bietti   Julien Mairal - Inria Grenoble

Understanding Deep Convolutional Representations

Are they stable to deformations?
How can we achieve invariance to transformation groups?
Do they preserve signal information?
How can we measure model complexity?

Kernel approach: construct functional space containing CNNs.

Why? Separate learning from representation: $f(x) = \langle f, \Phi(x) \rangle$

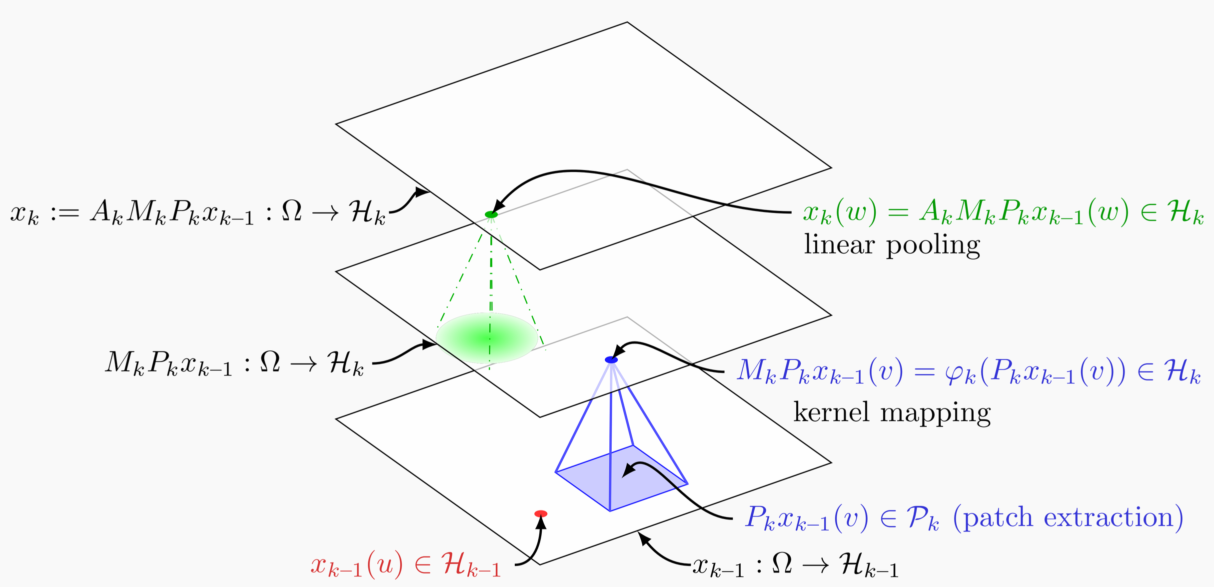
- $\Phi(x)$: CNN **architecture** (stability, invariance, signal preservation)
- f: CNN **model**, learning, generalization through $\|f\|$

$$|f(x) - f(x')| \leq \|f\| \cdot \|\Phi(x) - \Phi(x')\|$$

- $\|f\|$ **controls both stability and generalization**!

 $\rightarrow$ discriminating small deformations requires large $\|f\|$
 $\rightarrow$ learning stable functions is "easier"

Deep Convolutional Kernel Representation based on CKNs



- $x_0 : \Omega \to \mathcal{H}_0$: initial (**continuous**) signal
  - $u \in \Omega = \mathbb{R}^d$: location ($d = 2$ for images)
  - $x_0(u) \in \mathcal{H}_0$: value ($\mathcal{H}_0 = \mathbb{R}^3$ for RGB images)
- $x_k : \Omega \to \mathcal{H}_k$: *feature map* at layer k

$$x_k = A_k M_k P_k x_{k-1}$$

Patch extraction operator P_k.

Extract small patch of feature map x_{k-1} around each point u.

$$\|P_k x\| = \|x\|$$

Non-linear mapping operator M_k.

Pointwise non-linearity φ_k to each patch (kernel map).

$$\|M_k x\| \leq \|x\| \quad \text{and} \quad \|M_k x - M_k x'\| \leq \|x - x'\|$$

Holds for (tractable) **CKN approximations** by projection. [Mairal, 2016]
(Also holds for generic CNNs with spectral norm factor.)

Pooling operator A_k.

Linear Gaussian pooling at scale σ_k (typically exponential in k).

$$\|A_k x\| \leq \|x\|$$

Multilayer construction.

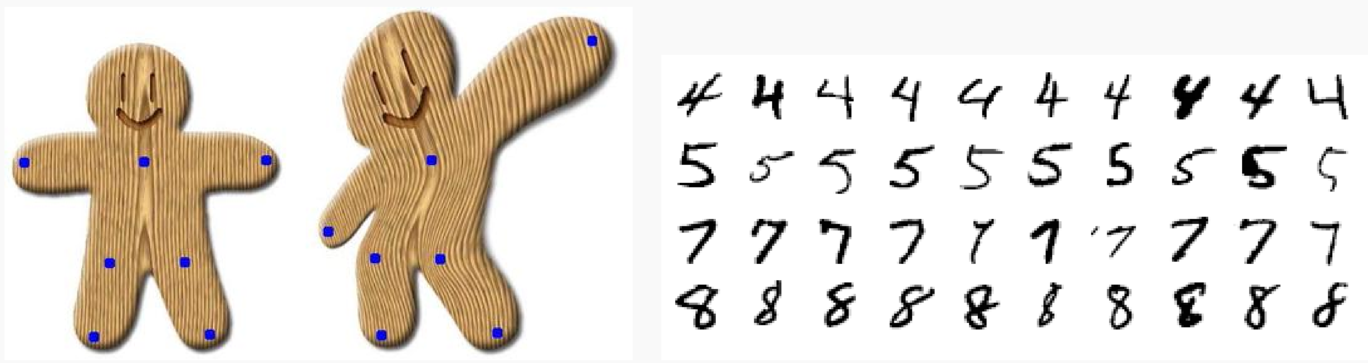
$$x_n := A_n M_n P_n A_{n-1} M_{n-1} P_{n-1} \cdots A_1 M_1 P_1 x_0 \ \in \ L^2(\Omega, \mathcal{H}_n)$$

Assume $x_0 = A_0 x$ (**anti-aliasing**), since x_0 is typically discrete

Invariance and Stability to Deformations

Deformations = ?

- $L_\tau x(u) = x(u - \tau(u))$: action of diffeomorphism $\tau : \Omega \to \Omega$
- Much richer group of transformations than translations



Definition of stability. [Mallat, 2012; Bruna and Mallat, 2013]

$\Phi(\cdot)$ is **stable to deformations** if

$$\|\Phi(L_\tau x) - \Phi(x)\| \leq (\underbrace{C\|\nabla\tau\|_\infty}_{\text{deformation}} + \underbrace{C'\|\tau\|_\infty}_{\text{translation}})\|x\|$$

Translation invariance. $L_c x(u) = x(u - c)$

- P_k, M_k, A_k commute with L_c: $\Box L_c = L_c \Box$

$$\|\Phi(L_c x) - \Phi(x)\| = \|L_c \Phi(x) - \Phi(x)\| \leq \|L_c A_n - A_n\| \cdot \|x\|$$

- Mallat [2012]: $\|L_\tau A_n - A_n\| \leq \frac{C_2}{\sigma_n}\|\tau\|_\infty$
- **Group invariance**: have P_k, A_k commute with $L_g x(u) = x(g^{-1}u)$
  - similar to [Cohen and Welling, 2016]
  - only need global pooling at last layer for global invariance

Stability to deformations.

- P_k and A_k do not commute with L_τ $\rightarrow$ study commutator $[\Box, L_\tau]$
- $[P_k, L_\tau]$ unstable at high frequencies $\rightarrow$ adapt to current resolution
- We show: if $\sup_{u \in S_k} |u| \leq \kappa\sigma_{k-1}$

$$\|[P_k A_{k-1}, L_\tau]\| \leq C_1 \|\nabla\tau\|_\infty$$

- C_1 grows as κ^{d+1} $\implies$ more stable with **small patches** (e.g. 3x3)

Theorem (**Stability**)

Let $\Phi_n(x) := A_n M_n P_n A_{n-1} M_{n-1} P_{n-1} \cdots A_1 M_1 P_1 A_0 x$.
If $\|\nabla\tau\|_\infty \leq 1/2$,

$$\|\Phi_n(L_\tau x) - \Phi_n(x)\| \leq \left(C_1 (1+n) \|\nabla\tau\|_\infty + \frac{C_2}{\sigma_n}\|\tau\|_\infty\right) \|x\|$$

Controlling stability.

- Full kernel method: $\|f\|_{\mathcal{H}_\mathcal{K}}$ (regularizer)
- CKN: $\|w_{n+1}\|_2$, ℓ_2 norm of last layer (regularizer)
- CNN: $\|w_{n+1}\|_2 \cdot \Pi_k \|W_k\|_2$ (??)

Signal Preservation for Kernel Representation

- Signal is preserved if discretized with subsampling $\leq$ patch size
- Recovery via linear measurements (need full kernel representation)
- For CKNs: depends on quality of kernel approximations

Model Complexity of CNNs

- RKHS contains CNNs with smooth homogeneous activations.
- RKHS norm controls generalization (complexity) and stability.

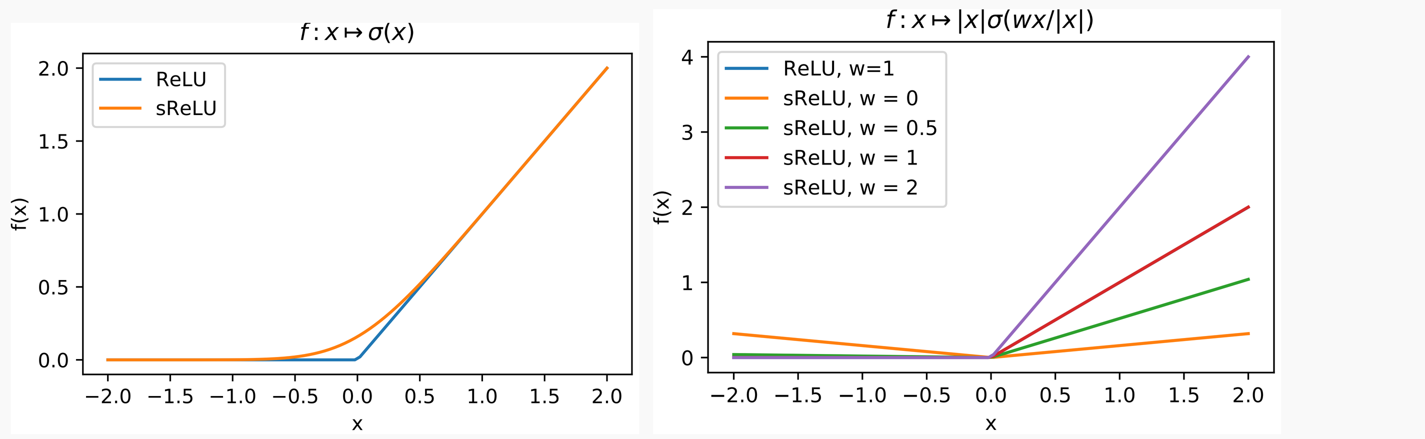
Patch kernels and their RKHS.

$$K_k(z, z') = \|z\|\|z'\|\kappa_k\left(\frac{\langle z, z'\rangle}{\|z\|\|z'\|}\right), \qquad \kappa_k(1) = 1$$

- e.g. Gaussian on sphere, inverse polynomial, etc.
- RKHS $\mathcal{H}_k$ contains, for smooth σ s.t. $C_\sigma^2(\|g\|^2) < \infty$,

$$f_g : z \mapsto \|z\|\sigma(\langle g, x\rangle / \|x\|) \qquad (\star)$$

- $\|f_g\|^2 \leq C_\sigma^2(\|g\|^2)$
- e.g. linear, polynomial, smooth ReLU
- Homogeneous version of [Zhang et al., 2016, 2017]



Construction of a CNN in the final RKHS.

- CNN f_σ with smooth homogeneous activations
- p_k feature maps at layer k, filters $w_k^{ij}(u)$, $W_k(u) := [w_k^{ij}(u)]_{ij}$
- Define intermediate ($\star$) functions (one per feature map)

Theorem (**RKHS norm of CNNs**)

The CNN function f_σ is in the RKHS $\mathcal{H}_\mathcal{K}$, with norm

$$\|f_\sigma\|^2 \leq p_n \sum_{i=1}^{p_n} \|w_{n+1}^i\|_2^2 B_{n,i},$$

where $B_{1,i} = C_\sigma^2(\|w_1^i\|_2^2)$ and $B_{k,i} = C_\sigma^2\left(p_{k-1}\sum_{j=1}^{p_{k-1}} \|w_k^{ij}\|_2^2 B_{k-1,j}\right)$.

Theorem (**RKHS norm using spectral norms**)

The CNN function f_σ is in the RKHS $\mathcal{H}_\mathcal{K}$, with norm

$$\|f_\sigma\|^2 \leq \|w_{n+1}\|^2 \, C_\sigma^2(\|W_n\|_2^2 \dots C_\sigma^2(\|W_2\|_2^2 \, C_\sigma^2(\|W_1\|_F^2)) \dots)$$

$\rightarrow$ **generalization** with Rademacher complexity and margin bounds.

Relevant References

- S. Mallat (2012).
  Group invariant scattering.
- Y. Zhang, P. Liang, and M. J. Wainwright (2017).
  Convexified convolutional neural networks.
- P. Bartlett, D. J. Foster, and M. Telgarsky (2017).
  Spectrally-normalized margin bounds for neural networks.