Article Presentation
Mathematics of Deep Learning

Authors: René Vidal, Joan Bruna, Raja Giryes and Stefano Soatto

Juan Carrascal, Maria Perpiñán, Edward Soto, Mayra Vega

Universidad Nacional de Colombia

May 2023

Introduction

The paper review recent work that aims to provide a mathematical justification for several properties of deep networks, such as global optimality, geometric stability, and invariance of the learned representations.

The structure of the paper is:

- The problem of training deep networks and conditions for global optimality.
- The invariance and stability properties of CNN.
- Structural properties of deep networks.
- Information-theoretic properties of deep representations.

Global Optimality in Deep Learning

We now study the problem of learning the parameters $W = \{W^l\}_{l=1}^L$ of a deep network from N training examples (X, Y). The problem of learning the network weights W is formulated as the following optimization problem

$$\min_{\{W^l\}_{l=1}^L} \mathcal{L}(Y, \Phi(X, W^1, \dots, W^L)) + \lambda \Theta(W^1, \dots, W^L), \tag{1}$$

where $\mathcal{L}(Y, \Phi)$ is the loss function and Θ is the regularization term.

Key challenge: For most deep networks, $\mathcal{L}(Y, \Phi)$ is a convex function, however the map $\Phi(X, Y)$ is not due to the product of the W^l variables and the nonlinearities in of the activation functions.
Why is this a problem?

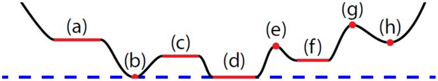

Fig 1: Example critical points of a non-convex function (shown in red). (a,c) Plateaus. (b,d) Global minima. (e,g) Local maxima. (f,h) Local minima.

Theoretical findings

In [1] they prove the following statements:

- For large-size networks, most local minima are equivalent and yield similar performance on a test set.
- The probability of finding a bad (high value) local minimum is non-zero for small-size networks and decreases quickly with network size.

Key challenge: Can we say something about the network optimality conditions without making any assumption on the input data distribution, the weight parameters or the network initialization?

Theoretical findings

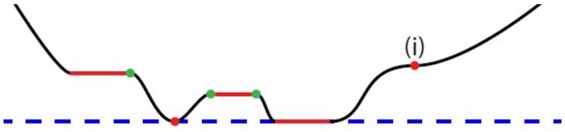

Fig 2: Critical points distribution [2, 3].

Theoretical findings

Theorem

If Φ and Θ are the sum of positively homogeneous functions of the same degree, then any local minimizer of the non-convex optimization problem

$$\min_{\{W^l\}_{l=1}^L} \mathcal{L}(Y, \Phi_r(X, W^1, \ldots, W^L)) + \lambda \sum_{i=1}^{r} \Theta(W^1, \ldots, W^L), \tag{2}$$

such that $(W_{i_0}^1, \ldots, W_{i_0}^L) = (0, \ldots, 0)$ for some $i_0 \in \{1, \ldots, r\}$ is a global minimizer of (2). Moreover, $X = \Phi_r(W^1, \ldots, W^L)$ is a global minimizer of (1).

Geometric Stability in Deep Learning

Convolutional architectures are the key to the success of most deep learning vision models.

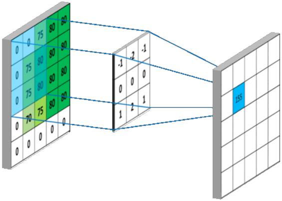

Fig 3: Convolutional filter

In these architectures, there is a notion of geometric stability which provides a possible framework to understand its success.

Geometric Stability in Deep Learning

Representation:
In image analysis applications, images can be thought of as functions on the unit square $\Omega = [0,1]^2$. This representation is given by: $X \in L^2(\Omega)$

$$X : \Omega \longrightarrow \mathbb{R} \;\text{ with }\; \int_{-\infty}^{\infty} |X(t)|^2 dt < \infty.$$

Let $f : L^2(\Omega) \longrightarrow \mathcal{Y}$ the unkown funtion we want to learn.

Stationarity:
Given a traslation operator

$$\mathcal{T}_v X(u) = X(u - v) \quad u, v \in \Omega, \tag{3}$$

we may consider following properties:

Invariance:
$f(\mathcal{T}_v X) = f(X)$ for any $X \in L^2(\Omega)$ and $v \in \Omega$. This is typically the case in object classification tasks.

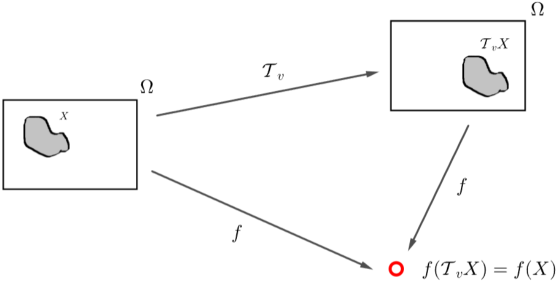

Fig 4: Invariance principle

Equivariance:
$f(\mathcal{T}_v X) = \mathcal{T}_v f(X)$ for any $X \in L^2(\Omega)$ and $v \in \Omega$. Observed in tasks of object localization and semantic segmentation.

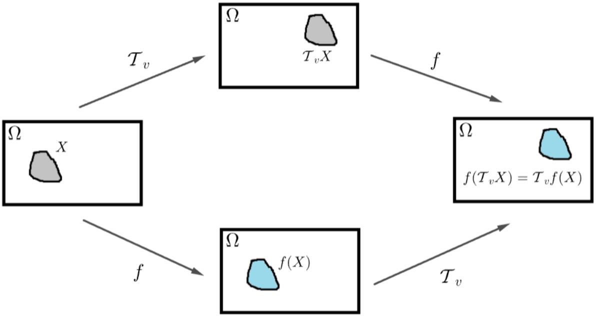

Fig 5: Equivariance principle

Local deformations:

Mathematically, given a smooth vector field $\tau : \Omega \longrightarrow \Omega$, if a function L_τ is such that:

$$L_\tau X(u) := X(u - \tau(u)), \tag{4}$$

we say that L_τ is a deformation. This kind of functions can model local translations, changes in viewpoint and rotations.

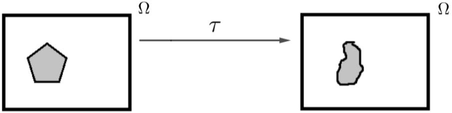

Fig 6: Local deformation

Geometric Stability in Deep Learning

Important facts:

- Most tasks studied in computer vision are not only translation invariant/equivariant, but, more importantly, also stable with respect to local deformations [4].
- Whereas long-range dependencies indeed exist in natural images and are critical to object recognition, they can be captured and down-sampled at different scales.
- CNNs strike a good balance in terms of approximation power, optimization, and invariance [5].
- Recently there has been an effort to extend the geometric stability priors to data that is not defined over an Euclidean domain [6].

Example:
Brain tumor segmentation task makes use of an architecture called *"encoder-decoder"*. We can observe the mentioned geometric properties in this experiment.

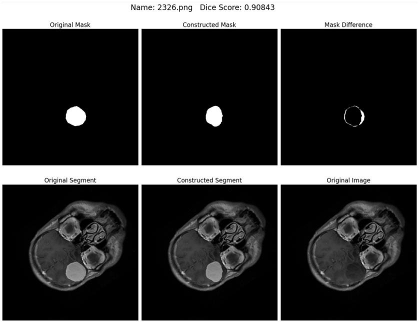

Fig 7: Brain tumor segmentation task

Key challenge: Understanding the good generalization observed in practice for deep networks with a large number of parameters or deep architectures.
The generalization error is then given as:

$$\mathrm{GE}(\Phi) = \left|\ell_{\mathrm{exp}}(\Phi) - \ell_{\mathrm{emp}}(\Phi)\right|.$$

Where:

$$\ell_{\mathrm{emp}}(\Phi) = \frac{1}{N}\sum_{X_i \in \Upsilon_N} \ell\left(Y_i, \Phi\left(X_i, W\right)\right),$$

$$\ell_{\mathrm{exp}}(\Phi) = \mathbb{E}_{(X,Y)\sim P}[\ell(Y, \Phi(X, W))].$$

Approaches: Measures such as **VC-dimension**, Rademacher or Gaussian complexities, and algorithm robustness have been used to bound the generalization error in deep networks.

Problem: Don't fully explain the good generalization observed in practice for DN with a large number of parameters or deep architectures.

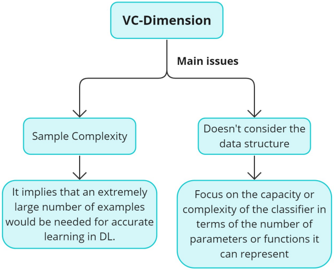

Fig 8: Issues with VC-dimension

Solution: New approaches such as **Classification margin**.

Definition

The classification margin of a training sample $s_i = (\mathbf{x}_i, y_i)$ measured by a metric d is defined as

$$\gamma^d(s_i) = \sup\{a : d(\mathbf{x}_i, \mathbf{x}) \leq a \Longrightarrow g(\mathbf{x}) = y_i \forall \mathbf{x}\}.$$

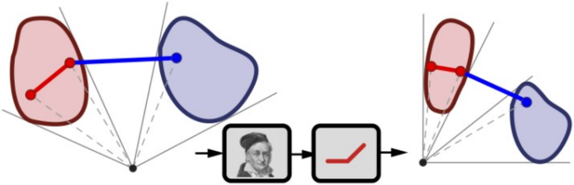

Fig 9: Sketch of the distortion of two classes with distinguishable angle

Solution: New approaches such as **Classification margin**.

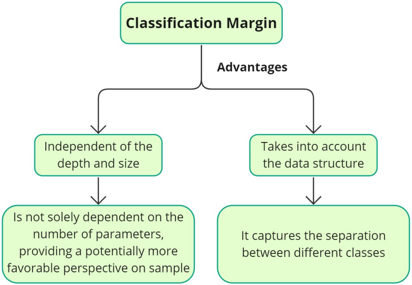

Fig 10: Advantages of classification margin

Towards an Information-Theoretic Framework

In multiclass clasification problems the loss function in neural networks is the *empirical cross-entropy*.

Definition

The *empirical cross-entropy*, as a loss function, is defined as:

$$\ell(W) = \mathbb{E}_{P(X,Y)}(-\log(\Phi(X, W)))$$

It is well known that to solve overfitting or high variance problems, techniques like regularization are implemented. The authors propose two different loss functions to address these problems.

Towards an Information-Theoretic Framework

From an information theory point of view, a type of regularization can be restrict the stored information in the training weights [7].

Train regularization

Suppose $P(W)$ is a prior on the weigts W. Define a loss by:

$$\ell(W) = H(Y|X, W) + \lambda KL(P(W|X, Y)||P(W))$$

Some problems were occurring in the attempt to calculate the Kullback-Leibler divergence. Recent advancements in the field of optimization have helped solve this problem.

Towards an Information-Theoretic Framework

It is well known that neural networks can adjust for noise in data. This idea can be explained as restricting the information extracted from the compressed representation of X by the network.

Test regularization

Suppose Z is a stochastic representation of X learned by a layer. Define a loss by:

$$\ell(W) = H(Y|Z, W) + \lambda I(Z; X)$$

Future work

Formally both approaches have no relation. But authors conjecture a relation exist and can be the explanation of generalization of networks. Bounds have been not found yet.

References I

[1] Anna Choromanska, Mikael Hena, Michael Mathieu, Gerard Ben Arous, and Yann LeCun.
The loss surfaces of multilayer networks.
International Conferenceon Artificial Intelligence and Statistics, pages 192,204, 2015.

[2] Benjamin D. Haeffele and René Vidal.
Global optimality in neural network training.
IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[3] Benjamin D. Haeffele and René Vidal.
Global optimality in tensor factorization, deep learning, and beyond.
arXiv, 2015.

[4] Stéphane Mallat.
Understanding deep convolutional networks.
Phil. Trans. R. Soc. A, 2016.

References II

[5] D. Ramanan P. Felzenszwalb R. Girshick, D. McAllester.
Object detection with discriminatively trained part-based models.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010.

[6] M. Bronstein, Y. LeCun J. Bruna, A. Szlam, and P. Vandergheynst.
Geometric deep learning: going beyond euclidean data.
arXiv preprint arXiv:1611.08097, 2016.

[7] Geoffrey E Hinton and Drew Van Camp.
Keeping the neural networks simple by minimizing the description length of the weights.
In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

[8] Marco Gori and Alberto Tesi.
On the problem of local minima in backpropagation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(1):76,86, 1992.

[9] Yoshua Bengio, Nicolas L. Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte.
Convex neural networks.
Neural InformationProcessing Systems, pages 123,130, 2005.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton.
Imagenet classification with deep convolutional neural networks.
Neural Information Processing Systems, pages 1097, 1105, 2012.

[11] Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues.
Generalization error of deep neural networks: Role of classification margin and data structure.
In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 147–151, 2017.

References IV

[12] Raja Giryes, Guillermo Sapiro, and Alex M. Bronstein.
Deep neural networks with random gaussian weights: A universal classification strategy?
IEEE Transactions on Signal Processing, 64(13):3444–3457, 2016.