



Choquet Capacity networks for random point process classification

Mehran Mohammadi, Santiago Velasco-Forero, François Willot, Mateus Sangalli, Jesus Angulo

► To cite this version:

Mehran Mohammadi, Santiago Velasco-Forero, François Willot, Mateus Sangalli, Jesus Angulo. Choquet Capacity networks for random point process classification. Proceedings of the 14th International Symposium on Continuum Models and Discrete Systems, Jun 2023, Paris, France. hal-04250560v1

HAL Id: hal-04250560

<https://hal.science/hal-04250560v1>

Submitted on 19 Oct 2023 (v1), last revised 9 Jan 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Choquet Capacity networks for random point process classification

Mehran Mohammadi, Santiago Velasco-Forero*, François Willot, Mateus Sangalli, Thomas Walter and Jesus Angulo

Mines Paris, PSL University, Center for Mathematical Morphology (CMM), France
Mines Paris, PSL University, Centre for Computational Biology (CBIO), France

Abstract. In this study, we revisit the Choquet capacity in the framework of convolutional neural networks, in $(\max, +)$ -algebra. By incorporating a discrete and learnable Choquet capacity model, we enhance the ability to represent the spatial arrangement and density variations in random point processes of convolutional neural networks. To validate the effectiveness of our approach, numerical experiments are conducted on synthetic datasets simulating diverse spatial point patterns of the Neyman-Scott process. When compared to classical convolutional neural networks, the proposed approach exhibits comparable or improved performances in terms of classification. Superior results are also observed in regression problems involving the Neyman-Scott parameter that monitors the point patterns spatial dispersion.

Keywords: Choquet capacity, Max-Plus Algebra, Deep Learning

1 Introduction

There is a significant demand for analytical techniques that can handle vast amounts of image data in particular in engineering science [3]. For instance, images from satellites, geological maps, microscope images of metals and materials, cellular tissue, and data from computerized tomography [7] all require automatic and quantitative methods due to the sheer volume of data involved. In some scenarios, the information observed is a set of coordinated in 2D or 3D, and the feature of interest depends on the spatial relationship between these points. For example, individual brain activation maps [36,21], RNA Localization patterns [11], positions on random sensor networks [22], realizations of random processes [20], among others. On one hand, mathematical morphology [26,29,19] offers some ideas that are applicable in this context, especially because the interpretation in term of size/shape relationship of interest object. On the other hand, data-driven learning methods have changed the way image problems are solved. Especially the deep models allow to solve recognition problems with performance sometimes superior to humans [15]. In this paper, we follow the results

* This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011012212R2 made by GENCI.

of [24,31,32,18], to present a new application of morphological models learned from data for the characterization of random point processes. The Neyman-Scott point process is presented in section 2. The Choquet Capacity model and the interpretation of the gradient descent learning mechanism is described in section 3. Section 4 demonstrate the benefits of the proposed model in comparison with classical convolutional neural networks for classification and regression problems on images generated by Neyman-Scott point process. We conclude in section 5.

2 Neyman-Scott point processes

In the present work, we consider Neyman-Scott point processes [34,33,35,23] as a way to generate point patterns that are statistically different from the classical Poisson model of points [6,5,17]. The Neyman-Scott model, like Poisson's, is a particular type of Boolean random set [30]. Poisson point processes on \mathbb{R}^2 depend on a single parameter θ , called the *intensity*. The intensity θ represents the mean number number of points per unit area. In the Poisson point model, the presence of points in two disjoint domains is uncorrelated which lead to the following Poisson-distributed law for the number of points $N(\Omega)$ in a domain Ω ,

$$\mathbb{P}\{N(\Omega) = n\} = \frac{(\theta|\Omega|)^n}{n!} \exp^{-\theta|\Omega|}, \quad (1)$$

where $|\Omega|$ is the surface area of Ω . This leads to the formula:

$$T_X(K) := \mathbb{P}(X \cap K \neq \emptyset) = 1 - \mathbb{P}\{K \subseteq X^c\} = 1 - \exp^{-\theta|K|}, \quad (2)$$

for the *Choquet capacity* $T(K)$ of a realization $X = \{x_1, x_2, \dots\}$ of a Poisson point process and a compact set K .

Neyman-Scott point processes are different from Poisson point processes in that they generate *clustered* sets of points. This is achieved in two steps. In the first step, a standard Poisson point process of intensity θ_p is generated, leading to a set of “parent points” $x^{(p)} \in \mathbb{R}^2$ ($p \in \mathbb{N}$). In the second step, we associate each parent point with a set of n_p “daughter points” $x_1^{(p)}, \dots, x_{n_p}^{(p)}$. The Neyman-Scott point process is made of all daughter points; parent points are not included. Neyman-Scott models are accordingly parametrized by θ_p , the probability law for the n_p and that of the daughter points, given in term of the multivariate probability law for the $x_i^{(p)} - x^{(p)}$. For simplicity, we consider a constant distribution $n_p \equiv n$. We also assume that the vectors $x_i^{(p)} - x^{(p)}$ are i.i.d. variables drawn from a multivariate random distribution in \mathbb{R}^2 , with uniform angular distribution in $[0; 2\pi]$. Distances from parent to daughter points are drawn from a Pareto law with parameters α, r_m such as:

$$\mathbb{P}\{|x_i^{(p)} - x^{(p)}| > r\} = \left(\frac{r_m}{r}\right)^\alpha H(r - r_m), \quad \alpha > 0, \quad r_m > 0, \quad (3)$$

where $H(\cdot)$ is the Heavyside function, equal to one on \mathbb{R}^+ and zero elsewhere. Here r_m controls the *typical size of the clusters* whereas the parameter α monitors the *spatial dispersion*. Thus, r has infinite variance when $\alpha \leq 2$ and infinite

mean when $\alpha \leq 1$. Examples of realizations are shown in Fig. (1). The points spatial dispersion in the Neyman-Scott model is markedly more heterogeneous when α is large (see Fig. (1b)). When α approaches one, the Neyman-Scott and Poisson models are difficult to distinguish (see Figs. 1).

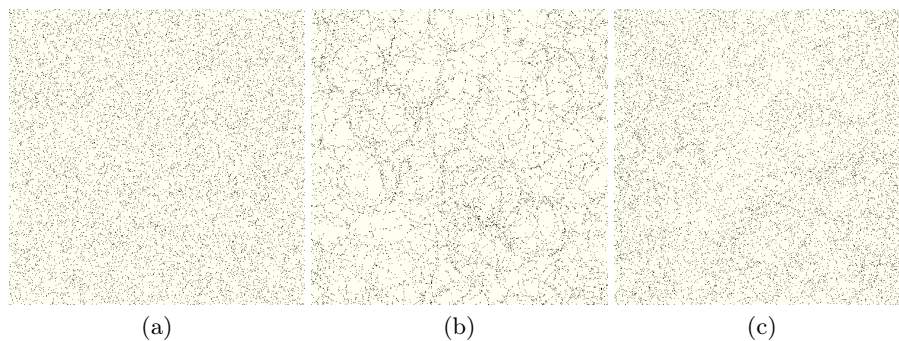


Fig. 1: Realizations of a Poisson point process (a) and Neyman-Scott models (b-c) with the same density of points. b) Parameters $r_m = 0.1$, $\alpha = 100$. c) $r_m = 0.1$, $\alpha = 3$.

Importantly, the Choquet-Matheron-Kendall theorem [4,14,16,30] states that any stationary and ergodic random set, such as Poisson or Neyman-Scott point processes, is completely determined by the functional $T(K)$. In the next section, we propose a model that use (2) for finding K in a discrete setting, from labeled data in a supervised classification problem.

3 Choquet Capacity models

3.1 Discrete random set

Usually random sets are defined in the sense of Euclidean geometry, as subsets of \mathbb{R}^d . However, in many applications, as in image processing, the domain is discrete, as pixels structures in 2D, $\Omega = \mathbb{Z}^2$. The discrete random set (DRS) X on \mathbb{Z}^2 is now defined by $X := \{(i, j) \in \mathbb{Z}^2 : x_{i,j} = 1\}$. The *discrete-capacity functional* of a discrete random set X is defined by [16]:

$$T_X(\mathbf{B}) := \mathbb{P}(X \cap \mathbf{B} \neq \emptyset) = 1 - \mathbb{P}\{\mathbf{B} \subseteq X^c\} := 1 - Q(\mathbf{B}) \quad (4)$$

for every $\mathbf{B} \in S$ the collection of all bounded subsets of \mathbb{Z}^2 . From (4) is easy to see that: a) $0 \leq T_X(\mathbf{B}) \leq 1$ for every $\mathbf{B} \in S$, b) $T_X(\emptyset) = 0$, and c) $T_X(\mathbf{B}_1) \leq T_X(\mathbf{B}_2)$, $\forall \mathbf{B}_1, \mathbf{B}_2 \in S$ such that $\mathbf{B}_1 \subseteq \mathbf{B}_2$. Other important properties can be shown for DRS [28] [9]. According to the Choquet-Kendall-Matheron theorem, any discrete random set X is uniquely characterized by its discrete-capacity

functional $T(\mathbf{B})$, defined over all bounded subsets $\mathbf{B} \in S$ of \mathbb{Z}^2 . Furthermore, we note that:

$$\mathbb{P}(X \cap \mathbf{B} \neq \emptyset) = \mathbb{P}(x \in \delta_{\mathbf{B}}(X)). \quad (5)$$

The capacity in (4) is accordingly given by the morphological dilation of the set X by \mathbf{B} [27,13].

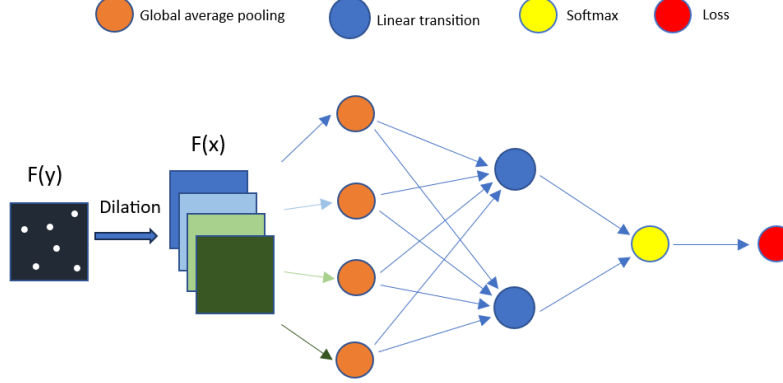


Fig. 2: Schematic of proposed Choquet Capacity model.

3.2 Choquet Capacity model

This section describes a model that can be used for the computation of (5), in the context of neural networks. To do so, we use the extension of convolutional models in the max-plus algebra[1], denoted “dilation layer” [31]. The dilation layer replaces classical convolutions and can be compared to fuzzy measures in so-called fuzzy integral neural networks [12], For an input image f^ℓ , the dilation layer is defined as:

$$\delta_i^\ell(x) := \bigvee_{y \in \mathbb{Z}^2} f^\ell(y) + \mathbf{B}_i(x - y) \quad (6)$$

where \mathbf{B}_i is the i -th kernel. In the cases where the image f^ℓ represents a DRS, and the kernel \mathbf{B} takes on values in $\{-1; 0\}^1$, we have that $f^\ell(y) + \mathbf{B}_i(x - y) = 0$ if $\mathbf{B}_i \subseteq f^\ell$ and 1 elsewhere. By sliding the kernel through the domain and calculating its average, assuming ergodicity, we obtain an estimate for the value of $Q(\mathbf{B}_i)$ in (4):

$$\eta_i^\ell := \frac{1}{m} \sum_{j=1}^m \delta_i^\ell(x_j) = \frac{1}{m} \sum_{j=1}^m \bigvee_{y \in \mathbb{Z}^2} [f^\ell(y) + \mathbf{B}_i(x - y)], \quad (7)$$

¹ Hereafter, we consider input images lying in the interval $[0, 1]$, so that in practice, -1 plays the same role as $-\infty$ as Kernel value.

where m is the number of pixels in the image. In the following, equation (7) is denoted *Choquet Capacity layer*.

Hereafter, we propose to study the performance of the Choquet Capacity layer to learn the optimal kernels \mathbf{B}_i in the case of supervised learning for the Neyman-Scott point process images introduced in section (2). The values computed in (7) are accordingly used as features of each input image, transformed linearly by means of a dense layer, and finally passed through the softmax activation layer:

$$q_j^\ell = \frac{\exp(\mathbf{W}_i^j \eta^i)}{\sum_{j=0}^1 \exp(\mathbf{W}_i^j \eta^i)}, \quad (8)$$

where W is the dense layer weight and $j = 1, 2$ denote the class number. Fig (2) shows a schematic of the proposed Choquet Capacity network. The classical training mechanism on neural networks seeks to find the optimal model's parameters W ,

$$\arg \min_{(\mathbf{B}, \mathbf{W})} \left\{ \sum_{i=0}^N \text{loss}(Y_i, \hat{Y}_i) + R(\mathbf{W}, \mathbf{B}) \right\} \quad (9)$$

where \hat{Y}_i is the prediction of the network for the i -th sample, Y_i is the ground truth, N is the number of samples and $R(\mathbf{W}, \mathbf{B})$ is a regularization term on the parameters of the model. In its simplest form, the parameters (\mathbf{W}, \mathbf{B}) are initialized to random values and updated by means of a stochastic gradient descend, the gradient being computed by backpropagation [15]. Note that the shape \mathbf{B} in (5) is a bounded subset, whereas the variable \mathbf{B} in (6) is used to compute a max-plus convolutional kernel with real values. Therefore, we propose two alternatives that address this difference during the training of the model:

- i) *Regularization*: the term R in (9) is set to penalize kernel values that are far from a binary equivalent, in our case 0 or -1 . Two different regularization functions are considered for R :

$$R_1(\mathbf{B}) = \lambda_1 \sum_{i=1}^{n_f} (\mathbf{B}_i + 1) + \lambda_2 \sum_{i=1}^{n_f} (\mathbf{B}_i + 1)^2, \quad (10a)$$

$$R_2(\mathbf{B}) = \lambda_1 \sum_{i=1}^{n_f} (\mathbf{B}_i + 1) \mathbf{B}_i + \lambda_2 \sum_{i=1}^{n_f} (\mathbf{B}_i + 1)^2 \mathbf{B}_i^2, \quad (10b)$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the trade-off hyperparameter to control the effect of regularization.

- ii) *Quantization* of the kernels during optimization. We use an integrated value filter to enforce kernel values in $[-1, 0]$. We use the “straight-through estimator” proposed in [2,37]. Predictions are computed by rounding the kernel to the closest integer (-1 or 0) whereas the gradient in the backpropagation is left unchanged.

Our Choquet capacity model is schematized in Fig. 2. The different approaches above are compared to one another and to classical convolutional networks in the next section.

4 Numerical experiments

In the present section, we quantify the performances of the various convolutional networks. Two problems are addressed: first, that of classification of realizations of Neyman-Scott point patterns w.r.t. Poisson; second, the prediction of the dispersion parameter α in a regression problem².

4.1 Classification of point pattern

In this experiment, we set the image size to 500×500 pixels and $\theta = 0.05$ in number of points per pixel unit. We choose $n = 50$ and let $\theta_p = \theta/n$ so that the density of points is the same irrespective of the model chosen. We also fix $r_m = 0.1L$ and let α varies between 1 and 100. All point patterns are statistically homogeneous. We perform the classification task on two classes of Poisson and Neyman-Scott point patterns with different values of the parameter α and conduct a comparative analysis with the Convolutions Neural Network (CNN). The data set contains 500 images for each class. For simplicity, we use one layer in each neural network, for the classification problem. Four filters with size 7×7 pixels are used.

For supervised classification, a common choice of loss function is the *binary cross-entropy*, defined by

$$\text{loss}(Y, \hat{Y}) = -\frac{1}{N} \sum_{\ell=1}^N \left[Y_{\ell} \log q_0^{\ell} + (1 - Y_{\ell}) \log q_1^{\ell} \right] \quad (11)$$

where (q_0^{ℓ}, q_1^{ℓ}) denote the prediction of the network for the l -th samples, N is the number of samples and Y_{ℓ} the ground truth.

The models are trained to predict the class between images produced by Poisson processes and Neyman-Scott models with different values of α parameter. Figure 3 shows the boxplot of validation accuracy for models training from different initialization values. Interestingly, the Choquet Capacity layer demonstrates competitive performance, especially when the parameter α decreases, so that realizations of the Neyman-Scott point pattern are similar to that of a Poisson distribution. Additionally, performance degrades when regularization or quantization is used. When regularization is applied, or quantization, the kernels obtained upon learning are close to a binary set, although no distinctive shape is observed (see Figure (4)).

4.2 Parameter regression of the Neyman-Scott point processes

In the present subsection, we compare the performances of different neural network models for predicting the value of the parameter α , in a regression problem. We set the image size to 128×128 pixels, the density of points to

² The source code of this section is available at <https://github.com/Jacobiano/ChoquetLayer>

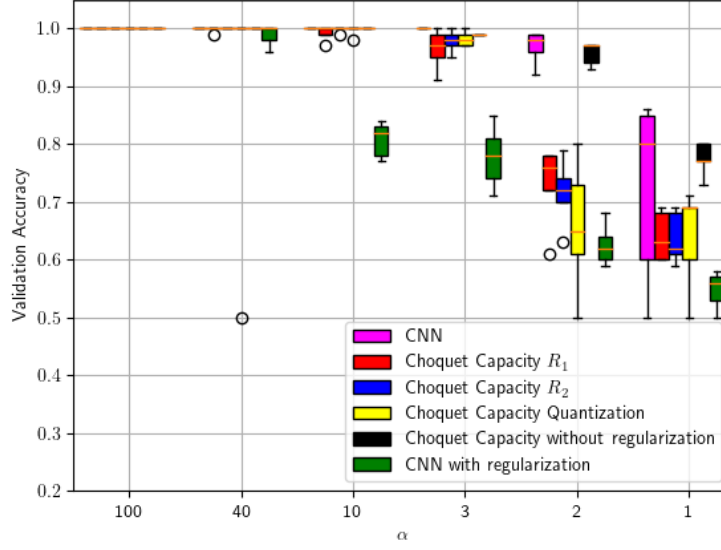


Fig. 3: Validation accuracy for the different neural network models in the two-classes classification problem (Poisson and Neyman-Scott point processes) vs. the parameter α that monitors spatial dispersion.

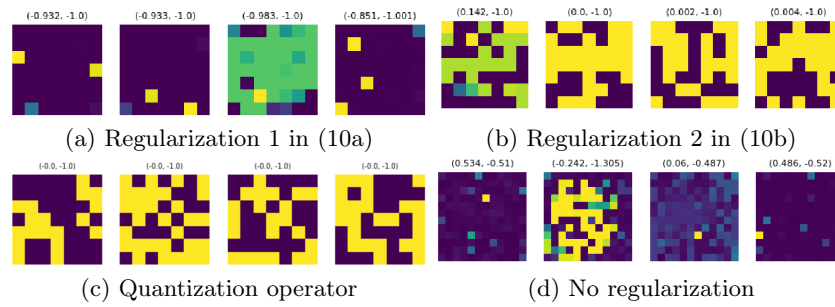


Fig. 4: Example of learned kernel \mathbf{B} using different proposed alternatives of the Choquet Capacity neural network.

$\theta = 100/(128 \times 128)$ and $r_m = 0.02$. We also let α vary between 0 and 9. This last is the value that we would like to predict. For the regression problem, the mean absolute value of the difference between true value and prediction:

$$\text{loss}(Y, \hat{Y}) = -\frac{1}{N} \sum_{\ell=1}^N |Y_{\ell} - \hat{Y}_{\ell}| \quad (12)$$

is used as loss function. During training the α values are scaled to the interval $[0, 1]$ as it is standard practice in regression models. As a first example, we compare the result between Choquet Capacity and a classical convolutional layer followed by a multi-layer perceptron (MLP) that consists of two layers of dimension ten and batch normalization [25], as well as a sigmoid activation, see Figure 5.

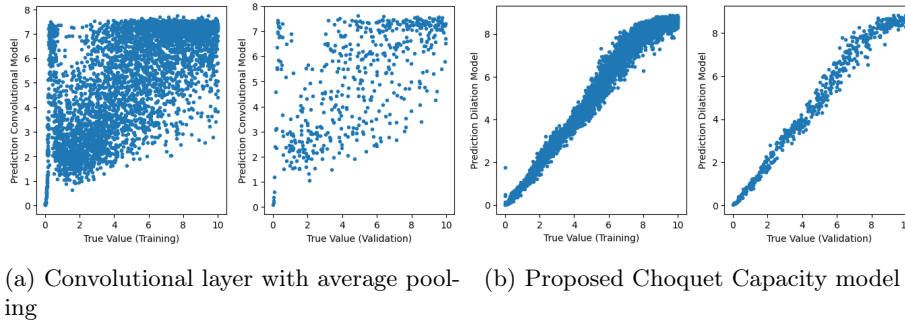


Fig. 5: Comparison on the prediction over testing training for one layer convolutional with global average pooling network and our Choquet Capacity proposition in the regression problem of the Neyman-Scott point processes. Both models have the same number of parameters and they have been trained with the same protocol.

However, we would like to study the effect of including more layers into the model. This is called the *depth* of the network in deep learning [8]. Usually, increasing both depth and width (number of filters per layer) helps to improve the model performance until the number of parameters becomes too high and stronger regularization is needed [10]. For that, we consider convolutional layers with kernel size of 3×3 pixels and 48 filters are concatenated before the use of global pooling and the MLP. This architecture is represented in Figure 6. We compare different convolutions in the last layer before the global pooling: max-plus (Choquet Capacity model), classical convolutional (CNN), and depth-wise convolutional (DWCNN). The latter is included because contains exactly the same number of parameters compared to our Choquet Capacity model. In Figure 7, we compare the validation loss of the three networks with the number of parameters in each model shown in logarithmic scale. The number of

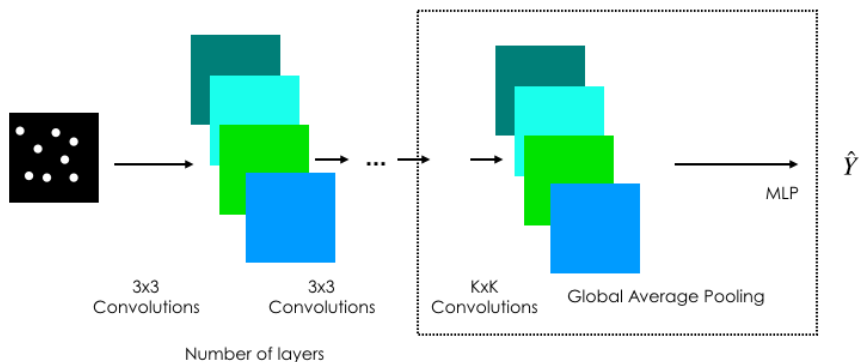


Fig. 6: In the architecture used in section 4.2, we vary the number of layers referring to classical convolutional layers of kernel size 3×3 pixels follow by ReLU activation.

degrees of freedom is increased by augmenting the depth from one to seven in the convolutional part in each architecture, leading to models with deeper feature extractions. The Choquet Capacity model improves the regression results in all explored cases, for a fixed number of parameters, compared to DWCNN. Comparable performances to the Choquet model are achieved by classical CNNs at the cost of a much higher number of parameters.

5 Conclusion

The results obtained in this study are based on a novel approach that makes use of dilation layers with global pooling, inspired from the definition of the Choquet capacity for random sets in the theory of integral geometry. Numerical experiments performed on synthetic datasets which simulate diverse Neyman-Scott point patterns, demonstrate the efficiency of the method, as well as seemingly superior performances in particular in regression problems involving the dispersion parameter in the Neyman-Scott point process. Compared to classical convolutions, the Choquet Capacity networks are somewhat more difficult to train, due to the presence of a nonlinear convolution filters. Yet they are versatile in that they may be combined with classical convolutional networks. As a negative result, the alternatives studied to motivate the estimation of binary kernels result in very important losses in the performance of the models. This topic, as well as the adequacy of Choquet capacity networks to analyse real-world data, will be addressed in future research works.

Acknowledgements The authors gratefully acknowledge the financial support of Institut Carnot (grant 220000496).

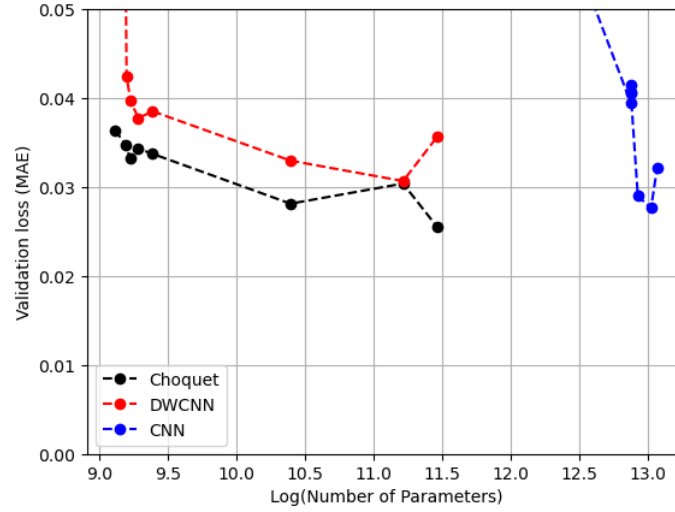


Fig. 7: Performance of different neural networks for varying number of layers in each model (See Figure 6). Three architectures are compared. In all cases each layer has 48 kernels of size 3×3 , ReLU is used as activation function, and a final layer with a convolution of size 13×13 followed by GlobalAveragePooling, BatchNormalization and a MLP with two layers of dimension ten with a final sigmoid activation. DWCNN refers to a depthwise convolution, CNN to a classical convolution and Choquet Capacity to the max-plus convolutions in 3. Note that DWCNN and Choquet Capacity models have exactly the same number of parameters.

References

1. Akian, M., Bapat, R., Gaubert, S., et al.: Max-plus algebra. Handbook of linear algebra **39** (2006)
2. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
3. Chinesta, F., Cueto, E., Abisset-Chavanne, E., Duval, J.L., Khaldi, F.E.: Virtual, digital and hybrid twins: a new paradigm in data-based engineering and engineered data. Archives of computational methods in engineering **27**, 105–134 (2020)
4. Choquet, G.: Theory of capacities. In: Annales de l’institut Fourier. vol. 5, pp. 131–295 (1954)
5. Daley, D.J., Vere-Jones, D.: An introduction to the theory of point processes: volume II: general theory and structure. Springer Science & Business Media (2007)
6. Daley, D., Vere-Jones, D.: An introduction to the theory of point processes: volume I: elementary theory and methods, Second Edition. Springer (2003)
7. Etienneble, A., Douillard, T., Valentin, O., Badot, J.C., Maire, E.: Numerical prediction of multiscale electronic conductivity of lithium-ion battery positive electrodes. Journal of The Electrochemical Society **166**(8), A1692 (2019)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
9. Goutsias, J.: Morphological analysis of discrete random shapes. Journal of Mathematical Imaging and Vision **2**, 193–215 (1992)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Imbert, A., Mueller, F., Walter, T.: Pointfish: Learning point cloud representations for rna localization patterns. In: European Conference on Computer Vision. pp. 487–502. Springer (2022)
12. Islam, M.A., Anderson, D.T., Pinar, A.J., Havens, T.C., Scott, G., Keller, J.M.: Enabling explainable fusion in deep learning with fuzzy integral neural networks. IEEE Transactions on Fuzzy Systems **28**(7), 1291–1300 (2019)
13. Jeulin, D.: Random texture models for material structures. Statistics and Computing **10**, 121–132 (2000)
14. Kendall, D.G.: Foundation of a theory of random sets. Stochastic geometry (1974)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
16. Matheron, G.: Random Sets and Integral Geometry. Probability and Statistics Series, Wiley (1974)
17. Møller, J., Waagepetersen, R.P.: Statistical inference and simulation for spatial point processes. CRC press (2003)
18. Mondal, R., Santra, S., Mukherjee, S.S., Chanda, B.: Morphological network: How far can we go with morphological neurons? In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press (2022)
19. Najman, L., Talbot, H.: Mathematical Morphology: From Theory to Applications. Wiley (2013)
20. Pawlasová, K., Dvořák, J.: Supervised nonparametric classification in the context of replicated point patterns. Image Analysis & Stereology **41**(2), 57–109 (2022)
21. Poline, J.B., Mazoyer, B.: Analysis of individual brain activation maps using hierarchical description and multiscale detection. IEEE transactions on medical imaging **13**(4), 702–710 (1994)

22. Ram, S.S., Manjunath, D., Iyer, S.K., Yogeshwaran, D.: On the path coverage properties of random sensor networks. *IEEE Transactions on Mobile Computing* **6**(5), 494–506 (2007)
23. Ripley, B.D.: Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(2), 172–192 (1977)
24. Sangalli, M., Blusseau, S., Velasco-Forero, S., Angulo, J.: Scale equivariant neural networks with morphological scale-spaces. In: *International Conference on Discrete Geometry and Mathematical Morphology*. pp. 483–495. Springer (2021)
25. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? *Advances in neural information processing systems* **31** (2018)
26. Serra, J.: *Image Analysis and Mathematical Morphology*. Academic Press Inc (1983)
27. Serra, J.: *Image Analysis and Mathematical Morphology: Vol.: 2: Theoretical Advances*. Academic Press (1988)
28. Singh, B., Siddiqi, M.: Discrete random functions: Modeling and analysis using mathematical morphology. *Mathematical Morphology and its Applications to Image and Signal Processing* pp. 65–72 (1996)
29. Soille, P.: *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, Berlin, Heidelberg, 2 edn. (2003)
30. Stoyan, D., Kendall, W.S., Mecke, J.: *Stochastic geometry and its applications*. John Wiley & Sons, Chichester (1995), second edition
31. Velasco-Forero, S., Pagès, R., Angulo, J.: Learnable empirical mode decomposition based on mathematical morphology. *SIAM Journal on Imaging Sciences* **15**(1), 23–44 (2022)
32. Velasco-Forero, S., Rhim, A., Angulo, J.: Fixed point layers for geodesic morphological operations. In: *British Machine Vision Conference* (2022)
33. Vere-Jones, D.: Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society: Series B (Methodological)* **32**(1), 1–45 (1970)
34. Waagepetersen, R.P.: An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics* **63**(1), 252–258 (2007)
35. Westcott, M.: Results in the asymptotic and equilibrium theory of poisson cluster processes. *Journal of Applied Probability* **10**(4), 807–823 (1973)
36. Worsley, K.J., Marrett, S., Neelin, P., Evans, A.: Searching scale space for activation in pet images. *Human brain mapping* **4**(1), 74–90 (1996)
37. Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., Xin, J.: Understanding straight-through estimator in training activation quantized neural nets. *ICLR* (2019)