Unsupervised learning of geometrical features from images by explicit group actions enforcement

Luca Bottero[1,3] Francesco Calisto[2,3] Valerio Pagliarino[1,3]

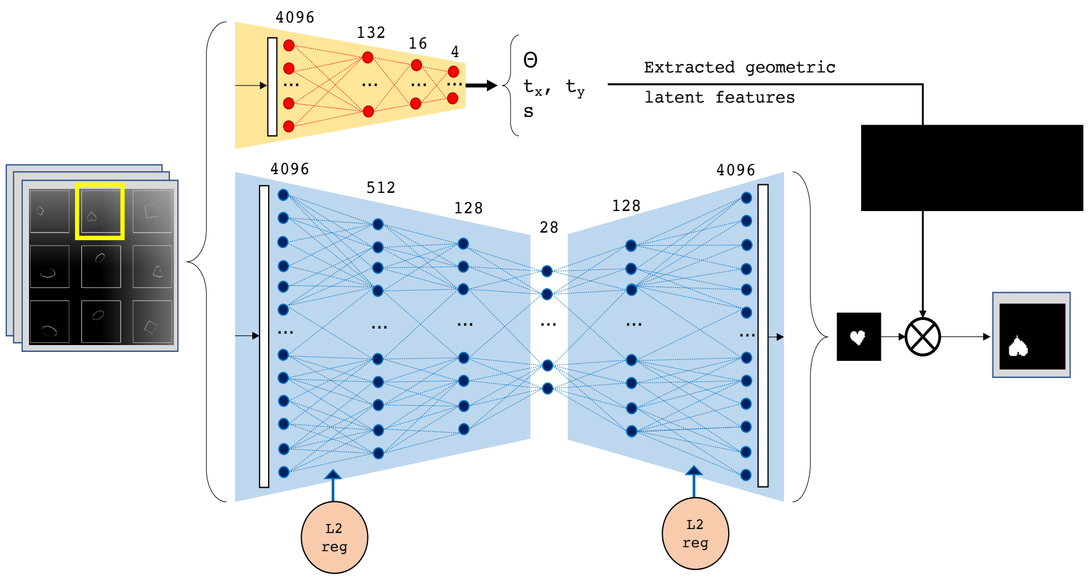
(1) University of Turin, Italy (2) LMU and TUM, Germany (3) MLJC, Italy

Introduction

Deep Neural Networks have widely proven to be an ideal tool to classify large datasets in various domains, computer vision being one of them. However, despite their numerous successes, sometime they lack generalisation and automatic meaningful feature extraction capabilities, which are effectively coherent with the task at hand. In addition, the leading paradigm for the training of NNs is supervised learning, which necessitates of costly labeling work.

Therefore, a recent direction of development in deep learning research is the investigation of new architectures possessing **invariance properties** with respect to specific transformations, endowing them with better generalization capabilities and a faster training process. A relevant example are architectures invariant with respect to some **geometric transformations**, encoded by a **group action**, which are able to process images by identifying objects independently of their location in space, similarly to what humans are able to do.

In this specific work, we aim at constructing a deep learning architecture with the ability to disentangle roto-translational and scaling properties of objects in 2D images from the ***intrinsic shape*** of the object in a fully automated way.



Architecture

Let's agree that by ***intrinsic*** or ***deep features*** S_i we mean the shape and topology of the objects inside the image, and by ***extrinsic*** or ***geometric features*** Θ_j (rotation θ, translation s, scaling t) the ones relative to their immersion in the 2D image space.
On the one hand, the ***geometric features*** are learned by an encoder with a 4-neuron dense output layer, corresponding to the object's rotation angle θ, scaling s and position coordinates (x,y). On the other hand, the ***deep*** ones are encoded in a 32-dimensional latent space and decoded to the original image by a dense deterministic autoencoder. The Θ_j parameters are used to operate an affine transformation on the decoded image, represented by the following matrix:

$$M = \begin{pmatrix} s\cos\theta & -s\sin\theta & t_x s\cos\theta - t_y s\sin\theta \\ s\sin\theta & s\cos\theta & t_x s\sin\theta + t_y s\cos\theta \\ 0 & 0 & 1 \end{pmatrix}$$

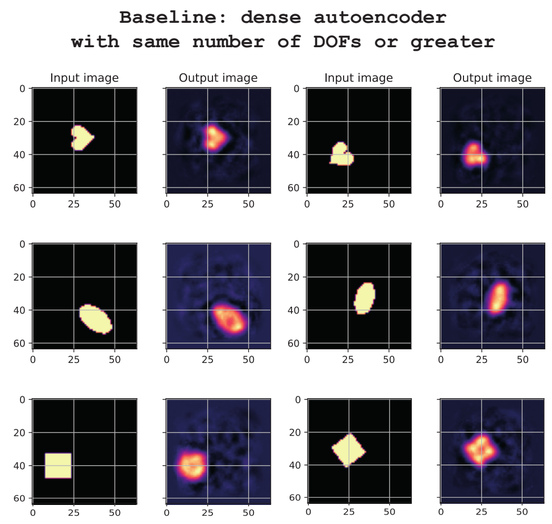
given by composition of rotations, translations and scaling matrices.
These represent affine invertible transformations from $V = \mathbf{R}^{64x64}$ to itself, i.e. they belong to the associated affine group $Aff(V)$. The latter are of the form $\Phi(a) = ga + v$, with $a, v \in V, g \in GL(V)$, which means $Aff(V) \cong V^{+} \rtimes GL(V)$.
Therefore Aff(V) is a Lie group. Being a composition of Lie group action matrices, M belongs to a Lie group itself.

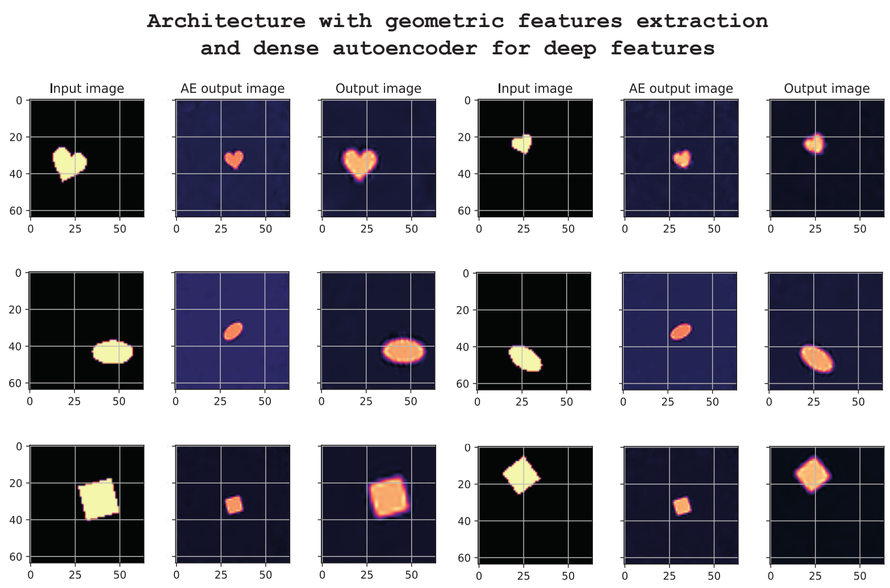
Training

The network was trained on a batch of 10000 images sorted randomly from the dataset and the mean squared error (MSE) was used as a loss function. The minimization strategy is a first-order gradient-based optimization with adaptive learning rate (ADAM algorithm), where the learning rate was initialised at 8e-4.
The ***dSprite*** dataset was chosen, as it contains images with clearly distinguishable shapes. It is a synthetically created dataset built combining the three kinds transformations (rotation, translation and dilatation) applied in sequence on three shapes (hearth, ellipsis, square).
Our architecture has been implemented using the ***PyTorch*** framework and trained using 10 x Intel Xeon cores, 20 GB RAM and an NVidia Tesla T4 GPU with 16 GB of memory. The model was run for 50 iterations and reached a final loss of 4.19e-3 in 6 minutes and 28 seconds.

Results

A few examples of our results on validation set are shown below, comparing the input image, the output of the dense autoencoder and the final output image.
On the right we include some results obtained with a classic dense auto-encoder with a larger number of degrees of freedom without any geometric enforcement: it is clear that the quality of the output image is considerably worse. It was trained in the same way and reached a loss of 7.58e-3 after 5:25 minutes.





Conclusions

We have devised a simple unsupervised architecture that was able to extract geometric features from images of 2D objects in an automated manner, by including a layer that enforces a group action. In the context of Geometric Deep Learning, we see this work going in a direction parallel to Physics-Informed Machine Learning, where physical dynamics are imposed, instead of the geometry: in both cases, making the best out of the a-priori-known structures is a powerful way to build more interpretable and efficient models, with generalization capabilities that elevate the process of learning to a more biologically-inspired endeavour.

Acknowledgements

A special thanks to Fondazione DIG421 and TesiSquare for their fundamental support towards our research work. We thank our colleagues Marco Nurisso and Luca Savant Aira for their contribution. FC acknowledges the TMP Program for the support. Furthermore, we are thankful to the University of Turin, HPC4AI, NPO Torino and Pompei StudentLab for their help.

NEURAL INFORMATION PROCESSING SYSTEMS

Dec 3rd, 2022, NeurReps Workshop, New Orleans, LA