SBGf
SBGf Conference
18-20 NOV | Rio'25
Sustainable Geophysics at the Service of Society
In a world of energy diversification and social justice

Submission code: KZVDQAXMZB

See this and other abstracts on our website: https://home.sbgf.org.br/Pages/resumos.php

A data augmentation workflow with geophysical concept-vectors

Julián L Gómez (YTEC/UNLP/CONICET), Emilio Camilion (Y-TEC S.A.), Danilo R Velis (CIGEOF/UNLP/CONICET)

SBGf Conference
18-20 NOV | Rio'25

A data augmentation workflow with geophysical concept-vectors

Introduction

Machine learning and deep learning offer unprecedented tools for digital signal processing, including geological and geophysical data. In these geosciences, datasets are most always expensive to obtain, proprietary, unbalanced, and limited. Generated samples from machine learning can lead to larger and balanced datasets to improve interpolation and classification tasks for reservoir characterization. The state-of-the-art in the generation of statistically consistent synthetic datasets comes from variational autoencoders (VAE), generative adversarial networks, and diffusion probabilistic models. The VAE family offer a latent space in which features of interest can vary smoothly and continuously. A typical VAE is a function composition of an encoder and a decoder. The encoder is a nonlinear geometrical transformation that compresses the input data into a latent vector space of reduced dimensionality. The decoder is another nonlinear function that connects the latent space with the larger ambient space of the input data. A VAE maps the unknown and complex density distribution of the data into a smooth distribution in the latent space. Sampling from the latent-space distribution results in novel synthetic samples in data space. The structure of the latent space created by the VAE allow us to explore specific directions, the concept vectors, where properties of interest can be manipulated to interpolate missing characteristics of the original dataset. A different approach is to train a conditioned VAE (CVAE), where the decoder is forced to produce samples that correspond to user-defined labels.

Method

We propose a workflow to manipulate the latent space of limited geological and geophysical datasets to generate samples conditioned on desired characteristics, such as roundness, and polarization light on images of segmented rock grains, and connectivity on rock permeability maps. The workflow includes checking the training stage with an attention-based mechanism to detect when the generated samples from the concept vectors can be considered as plausible new samples. We propose a novel loss that takes care on the reconstruction error by the concept vectors while considering the preservation of the original gray level of the training samples. To generate conditioned samples, a VAE can be trained in an unsupervised fashion and the concept vectors obtained by analyzing subsets of specific training samples and by optimizing their magnitude when translating samples from one subset into another. We show results from minimum hardware requirements. We select models which are simple enough to achieve the data-augmentation task with a few training epochs. Depending on the size of the input data, the VAE architecture is a dense neural network or a convolutional neural network.

Results and Conclusions

The results of our concept-vector data-augmentation workflow are promising: the interpolated datasets lead to more performant models of subsurface characterization. In contrast to CVAE, with concept vectors the properties of interest can be sampled at different degrees and added up. We find that generated samples via concept vectors can be more realistic in fewer training iterations.