Diffusion Model based Artificial Seismic Image Generation

Pedro Gil Oliveira de Magalhães Couto, Petrobras
Thales Amaral Paes de Mesentier, Petrobras
Vitor Giudice Batista de Araújo Porto, Petrobras
Ronnypetson Souza da Silva, Petrobras

This paper was prepared for presentation during the 4th Joint SBGf/SEG Workshop on Machine Learning held in Rio de Janeiro, Brazil, 03-04 December 2024.

Abstract

A challenge to using machine learning in the oil and gas industry is the cost of annotation. One common strategy to overcome this is the use of synthetic data. However, this usually leads to a distribution shift when applying the model to real data. This paper proposes using diffusion models to minimize the divergence in distributions, as these models progressively bring an image into the manifold of real data. Our aim is to understand if this technique can produce samples that (i) follow the statistical distribution of real data and (ii) contain qualitative characteristics that are not easily recreated by traditional methods. Our results show that the model can generate synthetic data that is indiscernible from real data to human discriminators – including experts in the field.

Introduction

The oil and gas (O&G) industry has been investing heavily in AI (Sharma, 2023), with predictions indicating continuous growth in the coming years. While being favored by solid investment capabilities and large amounts of raw data, O&G companies face challenges inherent to the development of machine learning applications. One of them is the lack of properly annotated data.

Traditional machine learning algorithms require labeled data to learn. Creating these labels, however, is generally a challenging task and demands large amounts of time and labor. Alternatives to circumvent the need for labeling data have been proposed, mainly using semi-supervised (Chapelle, 2006), unsupervised (Hinton and Sejnowski, 1999) or self-supervised (Sheng et al, 2023) machine learning algorithms or by generating synthetic data, with its corresponding labels, for traditional supervised training (Rombach et al, 2022).

Both approaches have their own set of limitations. Unsupervised, semi-supervised, and self-supervised learning have restricted use cases. The generation of synthetic data by traditional methods can be, depending on the complexity of the data to be reproduced, incredibly challenging and/or expensive and still lead to underwhelming results.

In the O&G industry, seismic acquisition imaging is one of the most difficult to label data types. Not only is the usual process of interpreting seismic data very demanding, but it is also uncertain. Certainty may be achieved only in association with other kinds of data – for example, well-boring profiles, which are rare and even more expensive than seismic data itself – and even so, it is somewhat restricted in scope. This scenario makes labeling seismic imaging very expensive and imprecise, making it a prime candidate for synthetic generation.

However, generating this kind of data is not simple. Properly reproducing the complex distributions and highly detailed qualitative characteristics of seismic imaging is very challenging and is hardly achieved by traditional techniques. The consequence is that the statistical distributions of artificial sets vary significantly from those of real data. Therefore, models trained on synthetic seismic data tend to have deteriorated performance when evaluated or tested over real data, partially defeating the technique's purpose.

But new techniques for generating images have been popularized in recent years. Amongst others, diffusion models have become state-of-the-art in image generation. Diffusion models are a class of machine learning models that generate data from pure Gaussian noise. They do that by learning a sequence of progressive denoising steps that bring images into the set of distributions belonging to the desired manifold (Rombach et al, 2022). Therefore, it is expected that they can generate hard to reproduce characteristics of real data. Recent works have explored diffusion models in applications using seismic data (Durall et al, 2023; Sian et al, 2023; Wei et al, 2023; Xiao et al, 2023; Trappolini et al, 2024; Zhang et al, 2024; Zhang et al, 2024)

This work proposes, then, the training of a diffusion model to unconditionally generate artificial seismic images and the evaluation of its performance. Our aim is to understand if the proposed technique can produce samples that (i) properly follow the statistical distribution of real data and (ii) contain qualitative characteristics that are not easily recreated by traditional computational and numerical methods.

Method

Our methodology consists of the following steps: (i) training the model; (ii) generating synthetic images; (iii) conducting the survey and (iv) statistically evaluating the results.

The proposed model is based on the unconditional image generation diffusion pipeline from the Hugging Face Diffusers library (Von Platen et al, 2022) and is composed of a DDPM (Ho at al, 2020) scheduler and a two-dimensional UNet (Ronneberger et al, 2015) denoising network.

The model is trained on an internal PETROBRAS dataset composed of 33,657 images extracted from 13 seismic volumes acquired over sedimentary basins of the Brazilian offshore subsea. Volumes are standardized, clipped to the (-5, 5) range and normalized to the (0, 1) range. During training, random crops with 256x256 resolution are taken from images and may be horizontally flipped with 0.5 probability. Due to the difficulty in properly representing the training objective via traditional metrics, evaluation of the model is mainly perceptual, with synthetic samples generated throughout training. Simple hyperparameter tuning is applied to further improve perceptual quality.

After model selection, 5,120 synthetic images are unconditionally generated. Another 5,120 are sampled randomly from the original dataset. This is the pool of images that is used in the survey. The hypothesis is that, in the absence of a clear objective metric, a good rule of thumb to assure the quality of the generated images would be to reach a point where subjects cannot distinguish them accurately from real images. In other words, the null hypothesis is that the expectation of correctly guessing if the provided sample is real or not will converge to 0.5 (the same as a random guess).

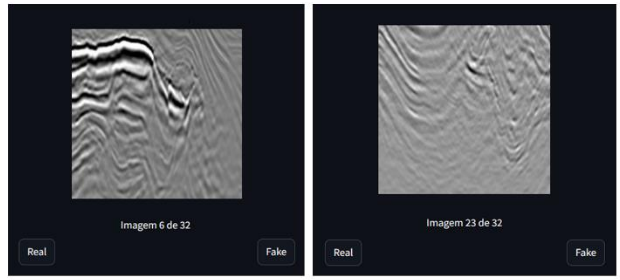
To evaluate this hypothesis, each respondent is provided with 32 images randomly sampled from the pool, with half of them being real, and the other half generated. This ensures that variance amongst individual responses is minimized. We also consider that any individual accuracy falling within two standard deviations from the expected mean to be merely a product of chance.

For conducting the survey, a simple web application was developed using python. When first accessing it, the users input their information data and obligatorily agree to the instructions. Then, for each person, a sequence of 32 images is sampled randomly from the pool, 16 of them being real, and 16 artificial. They are presented one at a time, in randomized order. The participants are never told how many images of each kind compose the survey. Users can click one of two buttons to answer whether they think that the image being presented is real or synthetic. After the 32 answers are given, they are stored, identified by the inputted information data.

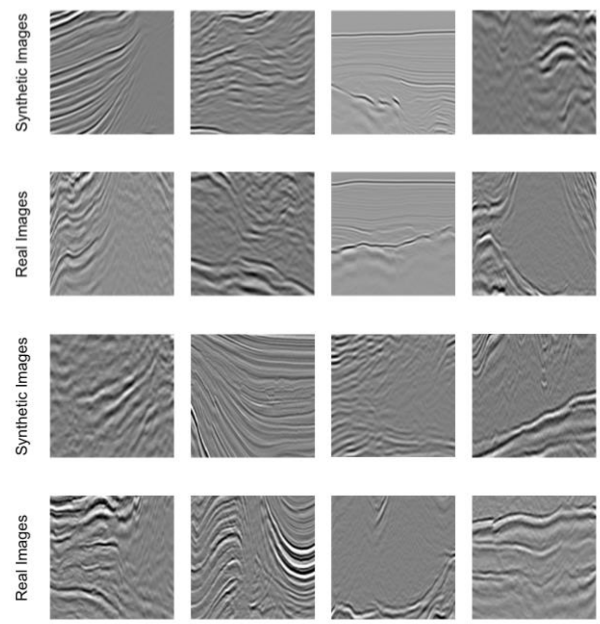
The selection of participants for the survey follows a set of constraints aiming to guarantee statistical relevance to the experiment. Firstly, all participants work at PETROBRAS. Secondly, groups of interest are defined to provide for ample coverage of O&G backgrounds and personal familiarity with seismic imaging, with a minimum number of participants from each of these groups being required. The groups are: (i) geologists, (ii) geophysicists and (iii) non-geoscientists – including engineers, developers, data scientists, designers, etc. The ranges of age, time working at the company and experience with geosciences vary freely.

Examples

The next figure illustrates the survey.



Below, a comparison between real and generated images, remarking the reproduction of complex characteristics, including geometry, texture, artifacts and noise, in the synthetic data.



Results

For this work, 81 participants have been surveyed. All the interest groups quotas have been met. Out of the 81 answers, 20 respondents are geologists, 44 are geophysicists and 17 fall into the non-geoscientist category.

The mean accuracy of all respondents is 49.7%, and the distribution of mean accuracy and standard deviation by group is:

	Geologists	Geophysicists	Non-geoscientists
Acc.	0.492	0.491	0.518
	(0.061)	(0.075)	(0.062)

The convergence of all interest groups' means and of the general mean towards 50% indicates that human discriminators have similar performance to a completely random approach, independently of their professional background or of their familiarity with seismic imaging.

This validates the hypothesis that the developed model can generate synthetic seismic images that are indiscernible from real ones to human discriminators – including experts in the field, indicating that the generated images reproduce complex distributions and qualitative characteristics of real data.

Qualitatively, the generated images do seem to reproduce patterns specific to seismic imaging. For example, the behavior of reflectors, assuring their continuity even in challenging geometries, and consistently representing varying intensities. Different kinds of geological structures are also mimicked in the synthetic set: ocean floor reflections, planar regions, folds, etc. Equally, acquisition effects such as ghosting, shadowed areas and other kinds of artifacts and noise that are of non-trivial artificial reproduction are present, as can be seen in the images in the example section.

Conclusions

The results of the survey show that diffusion models may generate synthetic seismic images that reproduce convincingly the statistical distribution and the complex characteristics of the real data – including geometry, texture, artifacts and noise.

The limited sizes of the generated samples and the following lack of larger structural elements to provide context don't allow for broader conclusions. The generation of entire seismic sections and volumes must be validated through future work. However, this work provides an important steppingstone in the way forward.

Acknowledgments

We would like to thank PETROBRAS and everyone who took part in the survey.

References

Chapelle, O. et al., 2006. Semi-Supervised Learning. The MIT Press. USA.

Durall, R., Ghanim, A., Fernandez, M., Ettrich, N., Keuper, J. 2023. Deep diffusion models for seismic processing. Computers & Geosciences, v. 177, 105377. doi:10.1016/j.cageo.2023.105377.

Hinton, G., Sejnowski, T. J., 1999. Unsupervised Learning: Foundations of Neural Computation. The MIT Press. USA.

Ho, J.; Jain, A.; Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. doi:10.48550/arXiv.2006.11239.

Hou, S., Wang, C., Zeng, Q., Cui, D., Yan, Y., Zhang, C., Zhang, Z., 2023. Seismic data interpolation based on diffusion model deep learning. In: Third international meeting for applied geoscience & energy expanded abstracts. DOI: 10.1190/image2023-3915640.1

Rombach, R., Blattman, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. doi:10.48550/arXiv.2112.10752.

Ronneberger, O.; Fischer, P.; Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. doi:10.48550/arXiv.1505.04597.

Sharma, G., 2023. How Multibillion Dollar Investments in AI are Driving Oil and Gas Sector Innovation. Forbes, https://www.forbes.com/sites/gauravsharma/2023/08/14/how-multibillion-dollar-investments-in-ai-are-driving-oil-and-gas-sector-innovation/. Accessed 25 September 2024.

Sheng, H., Wu, X., Si, X., Li, J., Zhang, S., Duan, X., 2023. Seismic Foundation Model (SFM): a new generation deep learning model in geophysics. doi:10.48550/arXiv.2309.02791.

Trappolini, D., Laurenti, L., Poggiali, G., Tinti, E., Galasso, F., Michelini, A., Marone, C., 2024. Cold Diffusion Model for Seismic Denoising. Journal of Geophysical Research: Machine Learning and Computation, v.1, n.2. doi:10.1029/2024JH000179.

Von Platen, P. et al, 2022. Diffusers: State-of-the-art diffusion models, https://github.com/huggingface/diffuse-rs. Accessed 25 September 2024.

Wei, X., Chang, C., Wang, H., Tan, C., Xiong, D., Jiang, B., Zhang, J., Kim, S., 2023. Seismic Data Interpolation via Denoising Diffusion Implicit Models with Coherence-corrected Resampling. doi:10.48550/arXiv.2307.04226.

Xiao, Y., Li, K., Dou, Y., Li, W., Yang, Z., Zhu, X., 2023. Diffusion models for multidimensional seismic noise attenuation and superresolution. Geophysics, v.89, n. 5. doi:10.1190/geo2023-0676.1.

Zhang, H., Li, Y., Huang, J., 2024. Conditional Denoising Diffusion Probabilistic Model for Seismic Diffraction Separation and Imaging, IEEE Transactions on Geoscience and Remote Sensing, v. 62, 5912413. DOI: 10.1109/TGRS.2024.3381193

Zhang, H., Liu, Y., Sun, Y., Chen, G., 2024. SeisResoDiff: Seismic resolution enhancement based on a diffusion model. Petroleum Science. DOI: 10.1016/j.petsci.2024.07.002