2025 EAGE ANNUAL 86TH CONFERENCE & EXHIBITION TOULOUSE | FRANCE

Advances in Scaling and Architecture of 3D Foundation Models for Seismic Data

T. Sansal[1], B. Lasscock[1], A. Valenciano[1]

[1] TGS

Summary

3D Seismic Foundation Models (SFMs) have been scaled to 1.8 billion parameters, pushing the boundaries of AI-driven seismic analysis. This work employs Vision Transformers (ViTs) augmented with multi-dimensional rotary positional embeddings and FlashAttention-2 to efficiently handle larger 3D spatial contexts. Pretraining was conducted on 20 terabytes of seismic data spanning 444,000 km² using a Masked Autoencoder (MAE) approach for self-supervised learning. Drawing on advancements in large model optimization, including key/query normalization and mixed precision techniques, the models achieved state-of-the-art generalization for salt segmentation tasks, with mean Intersection over Union (IoU) scores exceeding 0.9 across unseen datasets. Memory consumption analysis reveals a log-linear scaling relationship between model size, context size, and memory requirements. These advancements showcase the transformative potential of scaled SFMs in geophysical interpretation.

Advances in Scaling and Architecture of 3D Foundation Models for Seismic Data

Introduction

Foundation models are transforming artificial intelligence with their ability to perform various downstream tasks with minimal fine-tuning. These models learn generalized representations that transfer across tasks and domains by pretraining on massive datasets. While the idea of a foundation model has revolutionized natural language processing, computer vision, and multimodal AI, the seismic industry has yet to explore its potential.

Scaling laws provide a framework to estimate the relationship between dataset size, model size, and compute resources required for optimal performance in AI models. These laws are well established in natural language (Hoffmann et al., 2022) and natural images (Zhai et al., 2022). However, their application to scientific data remains largely unexplored. While Sheng et al. (2023), Lasscock et al. (2024), and Gao et al. (2024) have investigated seismic foundation models (SFM) and downstream tasks, these efforts were conducted at a relatively small scale.

This paper presents results from a large 1.8 billion-parameter 3D SFM trained on a large seismic dataset and discusses the practical considerations for training this model. We also address context size, which relates to the size of the 3D data the model can process. We provide details on how a large context model can be efficiently fine-tuned. The SFM generalization performance was benchmarked on the downstream task of salt segmentation on a hold-out dataset. The intersection over union (IoU) scores are compared to those of previous supervised studies by Roberts et al. (2024).

Dataset

The SFM was pre-trained using depth-migrated seismic data covering a surface area of 444,000 km^2 (20 terabytes size). We use non-overlapping 512^3 data cubes, which yield 54,000 unique seismic samples. This is equivalent to 1.8 billion unique visual tokens, each sized at 8x8x64 pixels in the inline, crossline, and depth dimensions. This dataset is intended to provide global geologic context, giving the pre-trained SFM the greatest capacity for generalization on various downstream tasks.

Method

We utilize the self-supervised learning approach of **M**asked **A**uto**E**ncoder (MAE) (He et al., 2022) and adapt to the seismic by tokenizing the data in the 3D image domain. **Vi**sion **T**ransformer (ViT) backbone provides simplicity, excellent scaling properties, and robust modeling capabilities. The MAE masking minimizes pre-training and fine-tuning memory requirements, allowing larger models and context sizes. The pre-training memory requirement was further reduced by using automatic mixed precision, data parallelism, and the level two **Ze**ro **R**edundancy **O**ptimizer (ZeRO) (Rajbhandari et al., 2020).

The first architectural modification is augmenting the tokenization and positional embeddings of the encoder/decoder models to process 3D data. Feichtenhofer et al. (2022) show that increasing the masking ratio can improve efficiency in spatiotemporal tasks. We follow the same approach and observe excellent reconstruction of 3D seismic images from 10% of the visible data (Figure 1). Second, we replace the positional embedding with **Ro**tary **P**ositional **E**mbedding (RoPE) (Su et al., 2024). Empirical studies show that RoPE performs equally or better than fixed sinusoidal positional embeddings. This is due to smoother positions in sequences achieved by applying positional encoding using a rotation matrix and the introduction of relative positional information. Our contribution here is to integrate multi-dimensional RoPE with the masking process of MAE and FlashAttention-2 (Dao, 2023). We use FlashAttention-2 to improve compute performance and reduce memory usage. Without FlashAttention-2 memory usage in training and inference would be a quadratic function of the number of input tokens. We apply position interpolation to RoPE to increase the context size via fine-tuning

(Chen et al., 2023). In addition to these significant architectural improvements, we also applied small architectural changes (Dehghani et al., 2023), like key/query normalization.

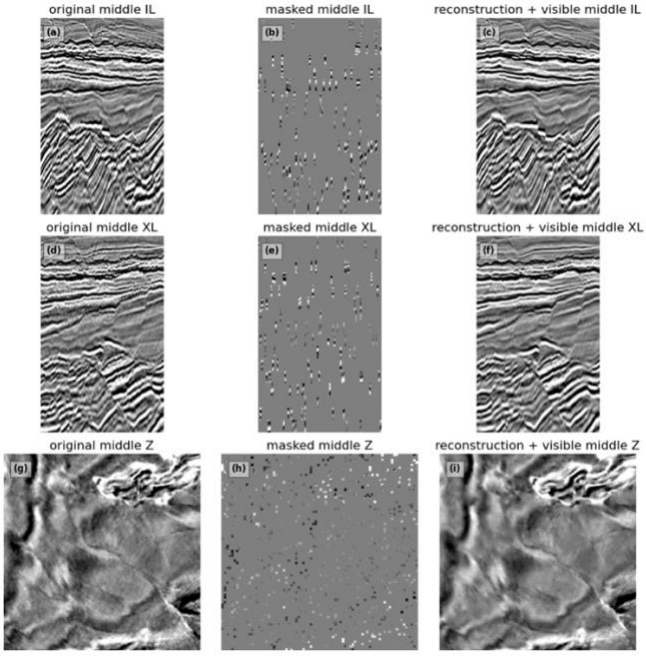

Figure 1. A sample of 640x640x1024 data and its reconstruction. (a-c) Mid-inline slice shows the original data, the data used input to reconstruction, and the reconstruction result. (d-f) and (g-i) show the equivalent crossline and depth slices, respectively.

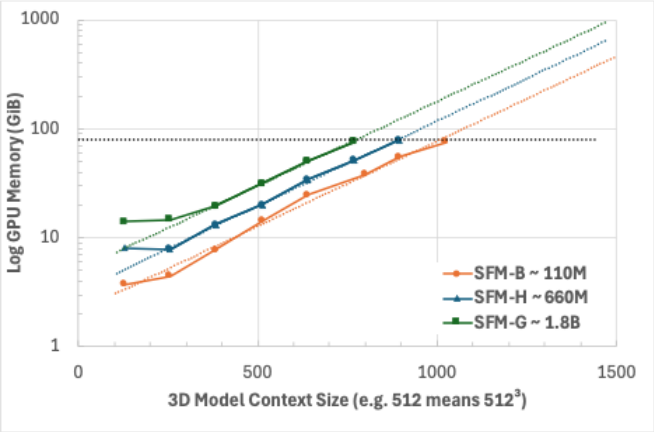

Figure 2. Memory consumption of SFM versus parameter count and context size.

Combining all the described transformer advancements, we can train memory and compute efficient 3D ViTs. We use the ViT naming and parameter convention: base (B), huge (H), and giant (G). SFM-B has 115 million, SFM-H has 660 million, and SFM-G has 1.8 billion trainable parameters. Figure 2 provides memory consumption as a function of model size and context size. When small contexts are excluded, we observe that the scaling relationship is log-linear. Memory consumption is significantly increased if we want to train very large contexts. Hence, we pre-train with smaller context and fine-tune to larger context. The figure extrapolates our measurements to even larger context sizes to estimate GPU resource requirements. The black dotted line represents a single A100/H100 memory limit (80GB).

Zhai et al. (2022) demonstrated that increasing dataset and model size improves vision models' task performance. Larger models consistently exhibit better few-shot learning capabilities, even when trained on smaller datasets. Our preference is for a larger model with a large context size. Based on this understanding, we decided to pre-train a 1.8 billion-parameter ViT-G model with a context size of 512^3

(32k tokens) and fine-tune it to an even larger 640x640x1024 cube (102k tokens). This study used a cluster of H100 GPUs and stored seismic data in MDIO format (Sansal et al., 2023) for efficient training and compute utilization.

Based on the findings of Roberts et al. (2024), we expect that a larger context enables segmenting larger objects more accurately, it reduces post-processing efforts and achieves more structurally continuous results. When training for this salt segmentation task we froze the encoder weights and initiated a new transformer decoder. We trained the salt segmentation models using the salt interpretation dataset previously utilized to train U-Net models in 2D and 3D, as described by Roberts et al. (2024) and Warren et al. (2023). The dataset comprises salt annotations from 23 RTM stacks in the Gulf of Mexico and South America for training. We added an interpreted RTM stack from South America for out-of-distribution testing and to measure the IoU metric evaluating the model's performance.

Results

The withheld stack from offshore South America was not included in pre-training and salt segmentation training to compare performance like our previous studies. The observed IoU values exceed 0.9 on the holdout dataset. Sections of the ground truth and predictions are illustrated in Figure 3.

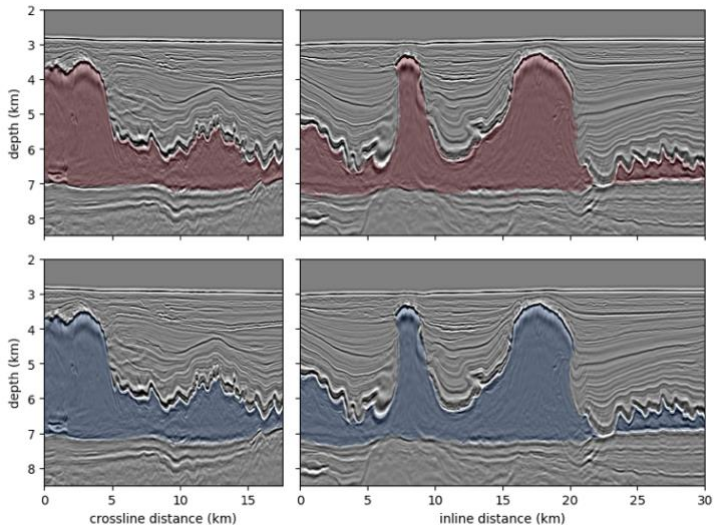

Figure 3. Seismic and masks for inline and crossline sections for the held-out dataset. The second row in blue shows the ground truth labels, while the first row in red shows the predictions.

Conclusions

Our study describes the intricacies of scaling the SFM to achieve state-of-the-art performance. It highlights the advancements that made it possible to pretrain a 1.8 billion parameter model, with a large context size, on a global seismic dataset. The MAE pre-training objective effectively reconstructs fine geological details from sparse inputs, showcasing its power in handling 3D seismic data. Achieving an IoU of 0.9 on held-out South American data, the salt segmentation task model demonstrates exceptional generalization to unseen data. The IoU performance aligns with state-of-the-art CNN-based supervised approaches with better generalization. Our advancements and insights in memory efficient training pave the way forward for scaling SFMs with more data and more parameters as AI training hardware improves.

Acknowledgments

We thank TGS for providing us with a world-class dataset and the appropriate computing resources, allowing us to pursue this research.

References

Chen, S., Wong, S., Chen, L. and Tian, Y. [2023] Extending context window of large language models via positional interpolation. *ArXiv*, 2306.15595

Dao, T. [2023]. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *ArXiv*, 2307.08691

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., et al. [2023] Scaling Vision Transformers to 22 Billion Parameters. Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, 202, 7480-7512.

Feichtenhofer, C., Fan, H., Li, Y., and He, K. [2022] Masked Autoencoders as Spatiotemporal Learners. Advances in Neural Information Processing Systems, 35, 35946–35958.

Gao, H., Wu, X., Liang, L., Sheng, H., Si, X., Hui, G. and Li, Y. [2024] A foundation model empowered by a multi-modal prompt engine for universal seismic geobody interpretation across surveys. *arXiv*, 2409.04962.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. [2022] Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, 15979-15988.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., et al. [2022] An empirical analysis of compute-optimal large language model training. Advances in Neural Information Processing Systems, 35, 30016-30030

Lasscock B., Sansal A., and Valenciano A. [2024], Encoding the subsurface in 3D with seismic. *SEG Technical Program Expanded Abstracts* : 617-621.

Rajbhandari, S., Rasley, J., Ruwase, O. and He, Y. [2020] ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. *arXiv*, 1910.02054

Roberts, M., Warren, C., Lasscock, B. and Valenciano, A. [2024]. A Comparative Study of the Application of 2D and 3D CNNs for Salt Segmentation. 85th EAGE Annual Conference & Exhibition, 1-5.

Sansal, A., Kainkaryam, S., Lasscock, B. and Valenciano, A. [2023] MDIO: Open-source format for multidimensional energy data. The Leading Edge, 42, 465–473.

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

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. [2024] RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568, 127063.

Warren, C., Kainkaryam, S., Lasscock, B., Sansal, A., Govindarajan, S. and Valenciano, A. [2023]. Toward generalized models for machine-learning-assisted salt interpretation in the Gulf of Mexico. The Leading Edge, 42(6), 390-398.

Zhai, X., Kolesnikov, A., Houlsby, N. and Beyer, L. [2022] Scaling Vision Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12104-12113.