

Non-contrastive approaches to similarity learning: positive examples are all you need

Alexander Marusov, Valerii Baianov, Alexey Zaytsev

Abstract—The similarity learning problem in the oil & gas industry aims to construct a model that estimates similarity between interval measurements for logging data. Previous attempts are mostly based on empirical rules, so our goal is to automate this process and exclude expensive and time-consuming expert labelling.

One of the approaches for similarity learning is self-supervised learning (SSL). In contrast to the supervised paradigm, this one requires little or no labels for the data. Thus, we can learn such models even if the data labelling is absent or scarce. Nowadays, most SSL approaches are contrastive and non-contrastive. However, due to possible wrong labelling of positive and negative samples, contrastive methods don't scale well with the number of objects. Non-contrastive methods don't rely on negative samples. Such approaches are actively used in the computer vision.

We introduce non-contrastive SSL for time series data. In particular, we build on top of BYOL and Barlow Twins methods that avoid using negative pairs and focus only on matching positive pairs. The crucial part of these methods is an augmentation strategy. Different augmentations of time series exist, while their effect on the performance can be both positive and negative. Our augmentation strategies and adaption for BYOL and Barlow Twins together allow us to achieve a higher quality ($\text{ARI} = 0.49$) than other self-supervised methods ($\text{ARI} = 0.34$ only), proving usefulness of the proposed non-contrastive self-supervised approach for the interval similarity problem and time series representation learning in general.

Index Terms—machine learning, deep learning, self-supervised learning, similarity learning, non-contrastive approaches, classification, interwell correlation

I. INTRODUCTION

The interwell correlation [1] is one of the important problems in the oil & gas industry. The target of this task is to understand how particular intervals in wells are similar to each other. An effective model of interwell correlation lets geologists plan oil production more effectively and detect additional risks of abnormal wells in advance. Moreover, hydrocarbon reserves estimation and building a geological field model need correct interwell correlation. However, changing sedimentation conditions across the basin make the process of interwell correlation rather tricky. Also, the manual interwell comparison is complex and tedious work [2]. Besides, the manual analysis is subjective because geologists can interpret the well-logging data differently.

Recently, more and more machine learning algorithms have been applied in the oil & gas domain [3], e.g. in rock type detection [4], [5]. Also, such kind of approach is used for

the similarity learning task. The authors of [6] propose using Overlap and Jaccard distances between well's representations received from the SVM algorithm. The paper [7] is based on the same idea but has another predictor. The next paper [8] follows classical supervised learning principles. The feature space is the aggregated statistics of time intervals, and the target shows whether well's intervals are similar or not.

Moving on to deep learning algorithms, the work [9] suggests using LSTM (Long Short-Term Memory) neural network to estimate correlation between production and injection wells. To exploit power of CNN (Convolutional Neural Network) with working with images the authors of the paper [10] propose to use deep CNN for working with well's images. The supervised paradigm of the approach still requires labels generated by experts. Authors of [1] consider different modern neural network architectures (Siamese, Triplet). They need no labels and work in a self-supervised regime. However, their models provide imperfect scores, suggesting that better approaches should exist that can capture more complex correlations and provide more universal representations.

Obtaining a universal description of an object is an extremely difficult but important task. SSL methods can be divided into two groups: generative or discriminative [11]. Discriminative methods consist from contrastive and non-contrastive methods. Contrastive methods use positive and negative pairs for training. One of the most famous frameworks for contrastive learning in computer vision, SimCLR [12], combines an additional projection head on top of the encoder output and contrastive loss. In the time series domain, authors of [13] propose to use a contrastive learning framework TS2Vec which provides a universal representation for any time series data. However, contrastive methods have some limitations. How it was mentioned above, one of the main problems of contrastive approaches is a complex labelling process. Also, these methods often need comparing each example with many others, which is time-consuming [14].

The recent works from the image domain use a non-contrastive paradigm. Non-contrastive methods use only positive pairs for training, usually obtained from augmentation of a single object, as we have no labels. The last mainly known self-supervised non-contrastive approaches in computer vision field are BYOL [14] and Barlow Twins [15]. BYOL uses two networks, where the student network tries to learn representations of the teacher network, and weights of the teacher network are exponential moving average of student network weights. It is important to notice that BYOL was also developed for audio data as well [16]. The idea of

A. Marusov, V. Baianov, A. Zaytsev are with Skolkovo Institute of Science and Technology (Skoltech), 121205. The work was supported by the Analytical center at Skoltech (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

Barlow Twins in the loss function which calculates as a cross-correlation matrix over embeddings of two identical networks. The concept of the method is to limit diagonal elements to one and non-diagonal elements to zero.

The methods from these works are based on learning to compare initial and augmented versions of the data, so selecting an augmentation strategy is the crucial point in model performance. There is a lot of research on image augmentations [17]. However, time series data augmentations coupled with deep learning received limited attention. The paper [18] compares different augmentation strategies on a time series classification task. The authors of [19] evaluate augmentations on several time series tasks like classification, anomaly detection and forecasting.

In this paper, we propose a deep learning model that considers the structure and properties of well-logging data. We use self-supervised methods because only unlabelled data is available for our task. We consider state-of-the-art non-contrastive approaches like BYOL and Barlow Twins, which do not require negative pairs and achieve a higher performance than contrastive algorithms.

The main contributions of the work are the following:

- 1) BYOL and Barlow Twins architectures were adapted for logging data from oil wells.
- 2) Different augmentation strategies were realized and compared.
- 3) To evaluate the quality of embeddings from BYOL and Barlow Twins, we trained different classifiers on top of them.
- 4) Our implementations of BYOL and Barlow Twins with optimal augmentations beat most self-supervised methods with $ARI = 0.49$ and as well show superior performance in downstream classification problem of classification of well interval type.

II. DATA

A. Data overview

We use an open well-logging dataset from the Taranaki Basin of New Zealand. It is provided by the New Zealand Petroleum & Minerals Online Exploration Database [20], and the Petlab [21]. The dataset consists of measurement of 21 different features from about 400 wells. They are standard logs for most geophysical companies that make transfer learning easy for wells of other formations. In our work, we use 4 features selected by experts, as other have a lot of missing values or unrelated for the similarity problem at hand. The description of them is provided in Table I.

Feature name	Description
DRHO	Porosity inferred from density
DENS	Density
GR	Gamma-ray
DTC	Sonic

TABLE I: Used features

We consider intervals from different wells and their similarity. Each interval mathematically represents as a matrix with dimension $\mathbb{R}^{l \times d}$, where

- l - interval length;
- d - number of features collected at each step of an interval.

We take the interval length equal to 100 similar to [1] and the number of features to 4.

B. Data preprocessing

Logging data from oil & gas wells have complex structure with missed data, high uncertainty and other peculiarities [22]. So, to construct a model on top of them we require careful preprocessing. It includes three consecutive steps: (a) filling missing values, (b) correcting sensor errors, and (c) normalization.

a) Filling missing values.: The oil & gas industry differs in the ratio of missing values. In our data, up to 70% of missing values are on each feature. Our strategy to fill missing values is the following: if we have an opportunity, we use *forward fill*, i.e. fill each missing value with the previously existing one. Otherwise - we use *backward fill*, i.e. fill with the closest future non-missing value.

b) Correcting sensor errors.: Another issue with the logging oil & gas data is a sensor error. Therefore, log outliers may be found in the data. The second step of data preprocessing is to drop physically inadequate data. Also, from the expert point of view, it is necessary to drop objects where the delta between CALI (calliper) and BS (bit size) is greater than 0.35.

c) Normalization.: The final data preprocessing step is normalization [23]. GR (gamma-ray) and NEUT (neutron), grouped by well and formation, are normalized by subtracting the mean and dividing to standard deviation, following [1]. Other features are normalized via the whole dataset.

III. METHODS

We start this section with the problem statement. Then we discuss, how can we adapt SSL paradigm for it and in particular BYOL and Barlow Twins approaches. We finish with discussing feasible augmentations. Also we include a separate subsection devoted to technical details to make the presented results reproducible.

A. Problem statement

To solve the interwell correlation task, we calculate the similarity between corresponding wells intervals. Consequently, we aim to create a method to output informative embedding for each interval. We expect embeddings of similar intervals to be closer in latent space than for different intervals. If we build such representations, we will have an opportunity to calculate similarity via various distances. One of the main approaches to build informative embedding is Self-Supervised learning (SSL).

B. SSL

Today most of the actual logging data is unlabeled [1]. SSL is a type of unsupervised learning where the pseudo labels are taken from the raw data itself. For example, in [24] authors

hide the word in sentence and predict it by its context. In our case, the dataset is unlabeled, and expert labelling is time-consuming and expensive. Instead we create pseudo labels for well similarity based on simple rules. This approach allows us to use self-supervised methods.

C. Non-contrastive SSL

Non-contrastive methods are based on representations of different views of one object. So representations of different views should be close to each other in latent space. In general, such an approach can lead to a collapse of representations, i.e. algorithm will produce constant output for all inputs. Each architecture solves collapse problem differently.

D. BYOL

The BYOL method [14] uses two networks: a student and a teacher network. The student network consists of an encoder, projector and predictor. The teacher network has the same modules as student except for predictor. There is no weight sharing between networks. However, the weights of the teacher network are the exponential moving average (EMA) of the student network weights, so we don't update the parameters for it in the usual way.

The main idea of BYOL is that the student network learns to predict the teacher network output. BYOL loss is defined as the mean squared error between the normalized prediction of the student network and the projection of the teacher network. Repeating this process iteratively, BYOL tries to improve representations. Combination of additional predictor in student network and idea of the moving average prevents collapse [14].

E. Barlow Twins

BYOL uses addition predictor and moving average to prevent collapse. But it is not clear how these components help to avoid collapse. To make SSL more understandable, the authors of Barlow Twins [15] suggest another approach.

Barlow Twins has a symmetric architecture. They forward two augmentations of an initial object to two encoders with similar architecture and similar parameters. Compared to BYOL, both networks are updated via a variant of gradient descent. Each network processes an augmented view of the object.

The loss is calculated over a cross-correlation matrix of representation outputs of both networks over a batch. The loss function consists of two terms. The first invariance term limits a diagonal of the cross-correlation matrix to one. It allows being consistent for input augmentations. The second redundancy reduction term pushes the sum of non-diagonal elements of the cross-correlation matrix to zero. It allows decorrelation of the components of representations.

F. Augmentations

Augmentation strategies significantly affect the final result of SSL. Considering the particular data at hand, we select jittering and window slicing as the augmentation strategies:

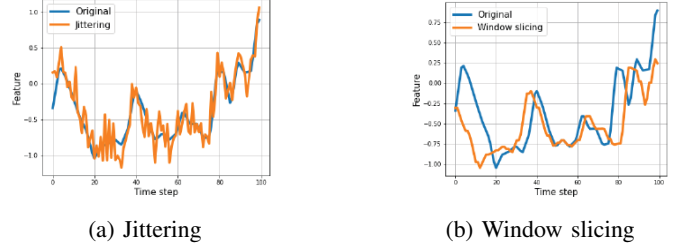


Fig. 1: Augmentations illustration

- 1) Jittering or adding an independent Gaussian noise at each point is one of the simplest augmentation strategies;
- 2) Random crop is one of the most potent image augmentations. For time series data, a window slicing is the term for the random crop.

In Fig. 1 influence of augmentation on feature time series can be seen.

G. Technical details

We use hyperparameters and procedures according to the best practices in the literature. The details of them are given below.

a) Encoder architecture: More and more tasks for time series data are solved by Recurrent Neural Network (RNN) [25]. RNNs often work better than more traditional techniques for Time series data [26]. Long Short-Term Memory or LSTM is a better variant of the basic RNN [27]. We use the LSTM unit as an encoder both for BYOL and Barlow Twins. The hidden size is 64.

b) Augmentations: We use jittering and window slicing as augmentation techniques. Also, an ablation study on searching best hyperparameters was carried out. For example, we varied the standard deviation for jittering, calculated along one feature from an interval or equal to 0.03, following the recommendations of [18]. In the second augmentation technique, the window size after slicing changes from 25 to 90, with the step equal to 5. We use random uniform selection of an interval for slicing.

c) Training protocol:

1) BYOL:

a) Architecture: Our projection and prediction heads have hidden and output dimensions 4096 and 256 correspondingly similar to [14]. The teacher network is the EMA of the student network weights with a momentum 0.99. The loss function is negative cosine similarity.

b) Optimization: The batch size is 64. We use a cosine annealing learning rate with a maximum number of iterations 10 and an initial learning rate is 0.1. During training, we use LARS [28] optimizer with momentum equal to zero. The duration of the training process was 20 epochs.

2) Barlow Twins:

a) Architecture: The projection head has the same architecture as [15] with the corresponding hidden and output dimensions 2048 and 2048. The loss function is the Barlow Twins loss [15].

b) *Optimization*: The batch size is 1024. The learning rate scheduler and optimizer are the same as in our BYOL implementation. The duration of the training process was 40 epochs. Following [15], the lambda parameter for the second term in the loss function is $5 \cdot 10^{-3}$

IV. RESULTS

We use clustering and classification evaluations to find the quality of self-supervised methods BYOL and Barlow Twins and compare them to the quality of baseline approaches. Before deepening to concrete results, let's first describe the general evaluation strategy.

A. Evaluation strategy

1) *Clustering*: Representations from BYOL and Barlow Twins may be used in various tasks. One of them is clustering. A model produces embeddings of the intervals. Similar intervals should have close embeddings and different intervals – distant embeddings. Hence, the idea is to cluster constructed representations and compare results to a selected ground truth labelling of intervals. We use a combination of class and layer expert's labelling as ground truth. More information about labelling is in [1]. As the clustering algorithm, we use agglomerative clustering that can work well for a relatively high embedding dimension.

Clustering evaluation can be conducted via several metrics. We use a standard Adjusted Rand Index (ARI) [29]. *ARI* ensures to have a value close to 0.0 for random labelling independently of the number of clusters and samples and exactly 1.0 when the partitions are identical (up to a permutation).

2) *Classification*: We also evaluate BYOL and Barlow Twins via several downstream classification tasks on top of obtained embeddings. The classifier model consists of two parts:

- *Encoder*. The encoder has a feature description of the interval on the **input** and fixed size embedding of this interval in the **output**.
- *Classifier*. Classifier receives embedding from the encoder on the **input** and predicts class at the **output**.

We freeze the encoder, and see, if a simple classifier can solve a problem on top of it. Thus, a better classifier score would suggest, that the embeddings are better representations.

We consider Binary classification for a pair of intervals and Multiclass classification for 28, 401 and 7 classes. The following is the formulation of the *Binary classification* task. The model considers a pair of intervals as *input*. The *output* should be 1 if these intervals belong to one well and 0 otherwise. The *Multiclass classification* model for a single *input* interval should predict which well the interval belongs to. There are 401 different wells. We also consider a smaller set of wells of size 28 selected by an expert. The *Geological classification* task is to predict the geological profile class for each well. The number of classes is 7. You can find more detailed formulations of the problems above in [1].

B. Evaluation results

In this section, we evaluate BYOL and Barlow Twins models. We selected hyperparameters using two strategies: maximizing ARI metric or accuracy in *Geological classification* task, see Section IV-C. Since we trained models with two different strategies, the names of BYOL and Barlow Twins have according clarifications in brackets. Our experiments show that both approaches have the same results for BYOL; consequently, there is no additional clarification for this architecture. We compare these architectures and Triplet [30] and VAE [31]. All models use a similar LSTM as an encoder.

As seen from Table II, the best model according to the ARI metric is BYOL, with all other methods being significantly worse.

Models	ARI
BYOL (ours)	0.49
Barlow Twins (ours, ARI)	0.32
Barlow Twins (ours, Geological)	0.26
Triplet	0.34
VAE	0.34

TABLE II: Comparison between models via clustering. The top value is in bold

Additional study is devoted to the examination of the classification results. For each classification task, we train different classifiers on top of the frozen embeddings: a logistic regression (Log.Reg.), a neural network with three fully connected network (FC-3) and Gradient boosting (XGBoost).

The results are in Table III. In *Binary classification* task, BYOL embeddings provide the best result. In *Multiclassification task (401 classes)* there is no sole winner. Triplet has slightly better results than other models on linear evaluation. BYOL shows the best result with the MLP classifier. Finally, Barlow Twins (Geological), selected according to geological classification, has the best result with XGBClassifier. In *Multiclassification task (28 classes)*, Triplet is the best one on linear and MLP evaluations, but Barlow Twins (Geological) again shows the perfect result with XGBClassifier. In *Geological classification (7 classes)* task, Barlow Twins (Geological) has the best performance on all evaluation protocols.

We note that our models beat the Triplet model on almost all tasks. So, depending on the solvable task, we can choose the best model according to Table III.

C. Hyperparameters selection

Selecting an appropriate augmentation strategy and optimization hyperparameters is more tricky than for supervised approaches. Since BYOL needs two distributions of image augmentations, we use window slicing as one augmentation technique and jittering as another. Unlike BYOL, the Barlow Twins approach needs augmentation from one augmentation distribution. Consequently, the first and second augmentations are identical.

1) Clusterization:

a) *BYOL*: A window size equals an almost maximum value of 85 and jittering, where deviation calculated via batch gives $ARI = 0.49$. Other values of hyperparameters provide comparable performances, with the lowest value being 0.41.

Models	Binary classification			Multiclass classification(401 classes)			Multiclass classification(28 classes)			Geological classification (7 classes)		
	Log.Reg.	FC-3	XGBoost	Log.Reg.	FC-3	XGBoost	Log.Reg.	FC-3	XGBoost	Log.Reg.	FC-3	XGBoost
BYOL (ours)	0.56	0.74	0.69	0.04	0.15	0.21	0.25	0.4	0.39	0.49	0.57	0.57
Barlow Twins (ours, ARI)	0.53	0.64	0.63	0.05	0.06	0.12	0.18	0.22	0.32	0.44	0.48	0.6
Barlow Twins (ours, Geological)	0.56	0.67	0.68	0.04	0.04	0.31	0.24	0.27	0.55	0.62	0.63	0.82
Triplet	0.55	0.69	0.68	0.08	0.11	0.23	0.4	0.44	0.46	0.59	0.61	0.59
VAE	0.54	0.69	0.68	0.03	0.06	0.15	0.19	0.29	0.32	0.45	0.54	0.55

TABLE III: Accuracy of different models for the classification problems

b) *Barlow Twins*: In Barlow Twins, the $ARI = 0.32$ is the best. The corresponding window size equals 50, significantly lower than the overall interval length of 100. The experiments using jittering as augmentation don't show comparable results.

2) Geological classification:

a) *BYOL*: How it was mentioned above, BYOL doesn't show any improvement comparing to the best BYOL model from clusterization task.

b) *Barlow Twins*: In Barlow Twins using window slicing with window size of 65 has the best results. You can see detailed comparison between two strategies of choosing the best model for Barlow Twins architecture in Table III.

V. CONCLUSION

We consider the challenging embedding construction problem for logging data from intervals in oil wells. We adapted self-supervised methods from the computer vision field: BYOL and Barlow Twins. According to our knowledge, we are the first who implemented these methods in oil&gas domain.

Since augmentation is a crucial part of these models, we made detailed ablation studies on searching best hyperparameters in augmentation.

We trained and evaluated these models with different strategies. In *Clusterization task* and in *Binary classification* BYOL shows the best results: $ARI = 0.49$ and corresponding accuracy results. In *Geological classification* problem, Barlow Twins (Geological) embeddings have the best accuracy with all classifiers on top of them. In rest *Multiclassification tasks* on half of the evaluation protocols, BYOL and Barlow Twins (Geological) show the best accuracy. Triplet provides top results in another half, but BYOL delivers comparable quality.

So, our non-contrastive approaches coupled with the right augmentation strategy provide better embeddings compared to other self-supervised methods. Thus, they are universal and suitable for the solution of diverse problems.

REFERENCES

- [1] E. Romanenkova, et al., Similarity learning for wells based on logging data, JPSE (2022).
- [2] A. K. Verma, et al., Assessment of similarity between well logs using synchronization measures, GRSL (2014).
- [3] D. T. dos Santos, et al., Deep recurrent neural networks approach to sedimentary facies classification using well logs, GRSL (2021).
- [4] E. Romanenkova, et al., Real-time data-driven detection of the rock-type alteration during a directional drilling, GRSL (2019).
- [5] A. Kadkhodaie-Ilkhchi, et al., Rock recognition from mwd data: a comparative study of boosting, neural networks, and fuzzy logic, GRSL (2010).
- [6] R. Akkurt, et al., Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction, SPWLA 59th Annual Logging Symposium (2018).
- [7] M. Ali, et al., Machine learning-a novel approach of well logs similarity based on synchronization measures to predict shear sonic logs, JPSE (2021).
- [8] A. Rogulina, et al., Similarity learning for well logs prediction using machine learning algorithms, in: IPTC, OnePetro, 2022.
- [9] H. Cheng, et al., LSTM based EFAST global sensitivity analysis for interwell connectivity evaluation using injection and production fluctuation data, IEEE Access (2020).
- [10] S. Brazell, et al., A machine-learning-based approach to assistive well-log correlation, Petrophysics-The SPWLA Journal of Formation Evaluation and Reservoir Description (2019).
- [11] S. Egorov, et al., Self-supervised similarity models based on well-logging data, arXiv preprint arXiv:2209.12444 (2022).
- [12] T. Chen, et al., A simple framework for contrastive learning of visual representations, ICML (2020) 1597–1607.
- [13] Z. Yue, et al., Ts2vec: Towards universal representation of time series, 2022.
- [14] J. Grill, et al., Bootstrap your own latent-a new approach to self-supervised learning, NeurIPS (2020).
- [15] J. Zbontar, et al., Barlow twins: Self-supervised learning via redundancy reduction, ICML (2021).
- [16] D. Niizumi, et al., Byol for audio: Self-supervised learning for general-purpose audio representation, IEEE IJCNN (2021).
- [17] D. Yarats, et al., Image augmentation is all you need: Regularizing deep reinforcement learning from pixels, ICLR (2020).
- [18] B. Iwana, S. Uchida, An empirical survey of data augmentation for time series classification with neural networks (2021).
- [19] O. Wen, et al., Time series data augmentation for deep learning: A survey, arXiv preprint arXiv:2002.12478 (2020).
- [20] Ibm research. the new zealand petroleum & minerals online exploration database (2015).
- [21] D. Strong, et al., Petlab: New zealand's national rock catalogue and geo-analytical database., New Zealand Journal of Geology and Geophysics (2016).
- [22] A. Acock, Working with missing values, JMF (2005).
- [23] J. Sola, J. Sevilla, Importance of input data normalization for the application of neural networks to complex industrial problems, IEEE Transactions on nuclear science (1997).
- [24] T. Mikolov, et al., Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2022).
- [25] D. Rumelhart, et al., Learning internal representations by error propagation, California Univ San Diego La Jolla Inst for Cognitive Science (1985).
- [26] H. Hewamalage, et al., Recurrent neural networks for time series forecasting: current status and future directions, IJF (2021).
- [27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation (1997).
- [28] Y. You, et al., Large batch training of convolutional networks, arXiv preprint arXiv:1708.03888 (2017).
- [29] W. Rand, Objective criteria for the evaluation of clustering methods, JASA (1971).
- [30] F. Schroff, et al., A unified embedding for face recognition and clustering, CVPR (2015) 815–823.
- [31] D. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).