TS2VecAR

Constantin von Crailsheim*
Department of Statistics
Ludwig-Maximilians-Universtität München c.crailsheim@campus.lmu.de

Abstract

.

1 Introduction

Time series data plays a significant role in science and industry in domains such as medicine, finance and manufacturing. However, Eldele et al. (2022) note that human annotation is very challenging, since time series patterns are not easily recognizable for humans, thus only litte time series data have been labelled (Ching et al., 2018). This indicates the need for self-supervised learning methods, which can learn the structure of the time series without labels in a pretext task and only needs few labelled instances to fine-tune the classifier.

Self-supervised representation learning proved in particular useful in computer vision with a contrastive loss function (too close to tstcc). For contrastive methods, invariant representations of the initial data are learnt by inducing similar representations for same instances but with different augmentations and different representation for different instances. Bachman et al. (2019) maximize mutual information between features (too similar) in multiple views to induce the algorithm to learn higher level characteristics of the data. Chen et al. (2020) proposed a simplified contrastive learning framework and showed the role of the composition of data augmentations.

Representations learning for time series has recently gained momentum. Franceschi et al. (2019) obtain a general-purpose representation by sampling positives as random sub-series and feeding them through a dilated convolution encoder, where the representation is evaluated with a triplet loss. Mohsenvand et al. (2020) extended the SimCLR framework (Chen et al., 2020) to a classification task for EEG time series data. Tonekaboni et al. (2021) propose a contrastive learning framework for non-stationary time series, where distribution of local signals should be distinguishable. Oord et al. (2018) use autoregressive models to predict future instances in the latent space, to induce a representations that captures relevant information about predicting future instances.

While previous work has induced latent representations, which fulfilled subseries consistency (Franceschi et al., 2019) and temporal consistency (Tonekaboni et al., 2021), Yue et al. (2021) argue that these strong assumptions may be violated in the case of level shifts and anomalies. Thus, they propose contextual consistency, which simply "treats the representations at the same timestamp in two augmented contexts as positive pair" (Yue et al., 2021, p. 8982). I used TS2Vec as a base for my model implementation. However, TS2Vec only evaluates the quality of the representation of each timestamp rather isolated and does not learn the structure of the time series in an autoregressive sense. Thus, I integrated the cross prediction task as proposed by Eldele et al. (2022) into the TS2Vec framework, which uses a context vector summarizing a sample of latent representation with a Transformer (Vaswani et al., 2017) as an autoregressive model to predict future latent representations.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

Thereby, with TS2VecAR I aim to derive robust contextual representations which also capture relevant information about future timestamps.

2 Method

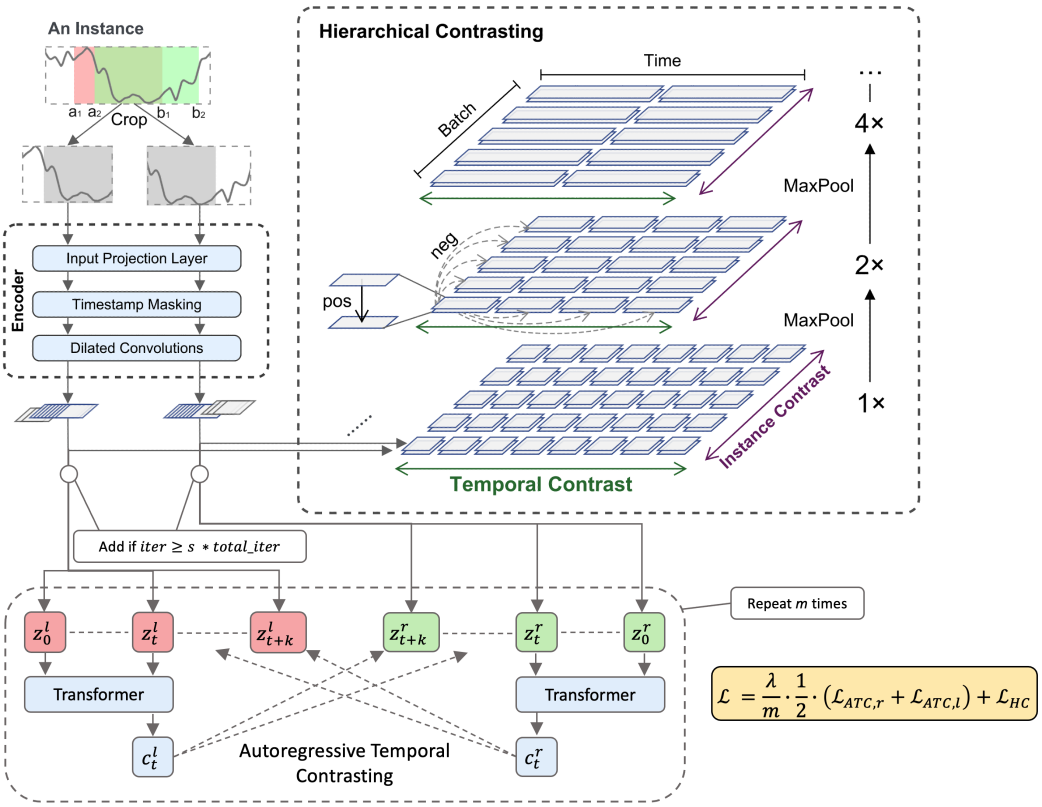

Figure 1: Structure of model (TS2Vec part copied from Yue et al. (2021) and ATC part is own illustration)

TS2VecAR integrates the temporal contrasting module of TS-TCC to the original implementation of TS2Vec. For TS2Vec, two overlapping windows of the time series are sampled, where the left window is defined by $[a_1, b_1]$ and the right window is defined by $[a_2, b_2]$. The following relationship has to hold for the cut-off points: $0 < a_1 \leq a_2 \leq b_1 \leq b_2 \leq T$. Then each window is encoded separately by feeding it through an input projection layer, timestamp masking and dilated convolutions. Subsequently, only the overlapping part (i.e. $[a_2, b_1]$) of the encoded sequence is kept. Each batch is then evaluated with hierarchical contrasting, which iteratively applies temporal and instance contrasting. Temporal contrasting takes the same timestamps from two representations of the same time series as positives and different timestamps from the same time series as negatives. Instance contrasting consider the two representations of the same time series as positives and the representations of different time series in the batch as negatives. By this the authors aim to achieve contextual consistency, which induces more robust learned representations. Refer to Yue et al. (2021) for more details.

However, the temporal contrasting only evaluates the quality of the representation of each timestamp rather isolated and does not learn anything about the structure of the time series in an autoregressive sense (see above). Thus, I integrated the idea by Eldele et al. (2022) to summarize the latent representation into a context vector by using an autoregressive model into the TS2Vec framework. To be precise, all latent representations up to sampled timestamp t (i.e. $\boldsymbol{z}_{<t}$) are passed to a Transformer,

which summarizes them into a context vector c_t. This context vector is used predict the next k latent representations as encoded from the other sampled window (cross prediction task). The prediction is evaluated with a contrastive loss (more details). Refer to Eldele et al. (2022) for more technical details.

The main adaptions are the following:

- While Eldele et al. (2022) use so-called strong and weak augmentations to generate representations of the time series in two different contexts, here two sampled and overlapping windows are used based on the TS2Vec model structure.
- The autoregressive model may only be included at later iterations to allow for the original TS2Vec part to already induce reasonable latent representations, which will be only refined by an autoregressive element at a later stage.
- Since the context vector is only used to predict a limited number of k subsequent latent representations in each iteration, the process of sampling t and cross predicting with a context vector can be repeated m times to exploit more of the time series.
- The final loss function is the sum of the hierarchical contrastive loss and the mean of the two autoregressive temporal contrastive losses, which were derived by cross predicting the latent representations from the left and right window.
- Additionally, a relative importance parameter λ, rescaled by the number of repetitions of the autoregressive model, is added.

3 Experiments

As downstream task, I used the classification of the entire time series as proposed by Yue et al. (2021) for comparability. Thus, a SVM classifier was fitted on the instance-level representation of the time series, which were derived by maxpooling over all individual timestamps. This allows to predict the classes of each instance and the performance was measured in accuracy.

For experiments, I used the subset of 12 UEA datasets to benchmark against TS2Vec, where Yue et al. (2021) showed SOTA performance in their paper. Since my implementation was done in a slightly different environment, I compared TS2VecAR to the replicated results of TS2Vec with their default settings as specified in their repository. The results of the experiments[2] are shown below, with different specifications of the share of iterations after which the autoregressive temporal contrasting was included.

Dataset	AR (s=0)	AR (s=0.1)	AR (s=0.2)	Replicated	Type
SelfRegulationSCP2	0.544	0.550	0.550	**0.556**	EEG
StandWalkJump	0.467	0.467	0.400	0.467	ECG
SpokenArabicDigits	0.967	0.959	0.977	**0.989**	Speech
DuckDuckGeese	0.460	0.460	**0.540**	0.520	Audio
ArticularyWordRecognition	**0.980**	0.970	0.977	0.977	Motion
CharacterTrajectories	0.991	0.993	**0.994**	0.992	Motion
EigenWorms	0.786	0.756	0.809	**0.863**	Motion
PenDigits	0.988	0.986	**0.990**	0.989	Motion
Handwriting	0.545	**0.551**	0.548	0.531	HAR
NATOPS	0.922	0.839	0.878	**0.939**	HAR
RacketSports	0.882	0.888	**0.895**	0.855	HAR
UWaveGestureLibrary	0.916	**0.925**	0.919	0.906	HAR
Mean (All datasets)	0.787	0.779	0.790	**0.798**	
Mean (HAR datasets)	**0.816**	0.801	0.810	0.808	

[2]As hyperparemeters I chose $\lambda = 5$ since the ATC loss is smaller than the HC loss and $m = 5$ to allow for sufficient cross prediction tasks in each iteration. All models were trained for the 600 iterations as specified default in TS2Vec.

In 7 out of 12 datasets, TS2Vec is outperformed by one specification of TS2VecAR. However, on average, TS2Vec has still the highest accuracy, which is partly due to the very low performance of the specification $s = 0$ and $s = 1$ on the DuckDuckGeese and Eigenworms dataset and of the specifiction $s = 0.2$ on the StandWalkJump and EigenWorms and NATOPS dataset.

Considering only the subsample of human activity recognition (HAR) datasets, TS2VecAR outperforms TS2Vec for three out of four datasets across all specifications. On average, the improvement is 0.8% when comparing the specification $s = 0$ to TS2Vec. However, the best individual performances are achieved by including the autoregressive temporal contrasting after some initial training solely with hierarchical contrasting. This suggests that learning a better representation initially before using it for the cross prediction task can be beneficial. However, their average accuracy is compromised by the low performance on the NATOPS dataset.

4 Conclusion

References

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *CoRR*, abs/1906.00910.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., and Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C.-K., Li, X., and Guan, C. (2022). Self-supervised contrastive representation learning for semi-supervised time-series classification.

Franceschi, J.-Y., Dieuleveut, A., and Jaggi, M. (2019). Unsupervised scalable representation learning for multivariate time series.

Mohsenvand, M. N., Izadi, M. R., and Maes, P. (2020). Contrastive representation learning for electroencephalogram classification. In Alsentzer, E., McDermott, M. B. A., Falck, F., Sarkar, S. K., Roy, S., and Hyland, S. L., editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 238–253. PMLR.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding.

Tonekaboni, S., Eytan, D., and Goldenberg, A. (2021). Unsupervised representation learning for time series with temporal neighborhood coding.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. (2021). TS2Vec: Towards universal representation of time series.

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.