← Go to **ICLR 2022 Conference** homepage (/group?id=ICLR.cc/2022/Conference)

Information-Aware Time Series Meta-Contrastive Learning

PDF **(/pdf?id=kxARp2zoqAk)**

Published: 29 Jan 2022, Last Modified: 14 Feb 2023 ICLR 2022 Submitted Readers: 🌐 Everyone Show Bibtex Show Revisions (/revisions?id=kxARp2zoqAk)

Keywords: Information-Aware Time Series Meta-Contrastive Learning

Abstract: Various contrastive learning approaches have been proposed in recent years and achieve significant empirical success. While effective and prevalent, contrastive learning has been less explored for time series data. A key component of contrastive learning is to select appropriate augmentations imposing some priors to construct feasible positive samples, such that an encoder can be trained to learn robust and discriminative representations. Unlike image and language domains where ``desired'' augmented samples can be generated with the rule of thumb guided by prefabricated human priors, the ad-hoc manual selection of time series augmentations is hindered by their diverse and human-unrecognizable temporal structures. How to find the desired augmentations of time series data that are meaningful for given contrastive learning tasks and datasets remains an open question. In this work, we address the problem by encouraging both high fidelity and variety based upon information theory. A theoretical analysis leads to the criteria for selecting feasible data augmentations. On top of that, we employ the meta-learning mechanism and propose an information-aware approach, InfoTS, that adaptively selects optimal time series augmentations for contrastive representation learning. The meta-learner and the encoder are jointly optimized in an end-to-end manner to avoid sub-optimal solutions. Experiments on various datasets show highly competitive performance with up to 11.4% reduction in MSE on the forecasting task and up to 2.8% relative improvement in accuracy on the classification task over the leading baselines.

One-sentence Summary: Guilded by information theory, we employ meta-learning mechanism to propose an information-aware time series contrastive learning approach that adaptively generates suitable task-specific augmentations for contrastive representation learning.

Supplementary Material: ⬇ zip (/attachment?id=kxARp2zoqAk&name=supplementary_material)

Add **Public Comment**

Reply Type: all Author: everybody Visible To: all readers **22 Replies**

Hidden From: nobody

[−] **Paper Decision**

ICLR 2022 Conference Program Chairs

21 Jan 2022 ICLR 2022 Conference Paper947 Decision Readers: 🌐 Everyone Show Revisions (/revisions?id=X_K91_oRpIb)

Decision: Reject

Comment: This paper presents a method which selects feasible data augmentations suitable for contrastive time series representation learning. The topic in this paper is timely and interesting. One of 4 reviewers did not complete the review, not responding to a few reminders. So, one emergency reviewer, who is an expert in meta-learning was added. While there is one review that strongly supports this work, two reviews remained unsupportive after the discussion period ended. I appreciate the authors for making efforts in responding to reviewers' comments. However, after the discussion period, most of reviewers had concerns in this work, pointing out that the technical correctness needs further justification and experiments should be improved. While the idea is interesting, the paper is not ready for the publication at the current stage. I encourage to resubmit the paper after addressing these concerns.

Add **Public Comment**

[−] Official Review of Paper947 by Reviewer ikmv

ICLR 2022 Conference Paper947 Reviewer ikmv

24 Nov 2021 ICLR 2022 Conference Paper947 Official Review Readers: 🌏 Everyone Show Revisions (/revisions?id=iU1W2OwuQSn)

Summary Of The Paper:
This paper proposes InfoTS, a method for learning augmentations that improve contrastive learning of time series data. The core contribution is a learnable augmentation strategy that uses Concrete/Gumbel-Softmax distributions. The paper shows empirical results on several time-series forecasting and classification benchmarks.

Main Review:
Concerns

- page 4, right after eq (3): "We dismiss the first part since the unconstrained entropy of v can be dominated by meaningless noise." Why is it reasonable to drop H(v)? This is a pretty big assumption you make early on, which is never addressed in the paper.
- Properties 1 and 2 are fairly trivial statements, which I don't think need proof.
- I think this paper is best described as an online hyperparameter tuning method rather than a meta-learning method. While one can argue that hyperparameter tuning is an instance of meta-learning, I felt that this naming is misleading in this context.
- All experiments: seeing error bars would be helpful because the differences between methods are fairly small in some cases.
- Table 3: these results are so close that I think many of them are not statistically significant. In particular, removing components from the meta objective seems to have minimal effect on final performance.
- I am not fully convinced that the proposed paper brings enough benefits to justify its complexity. For example, in Figure 8 and Table 5, the model mostly up-weights the subsequence policy, and the final performance is not very different from only using subsequence.

Minor comments

- Minor grammatical inaccuracies and typos throughout the paper. e.g., "the information theory," "important weight," "combing candidate," to name a few.
- page 3, "High Fidelity" paragraph: you call the augmented sequence v an augmentation. This is potentially confusing because augmentations typically refer to the mapping x->v, which you call g later.
- page 3, "High Fidelity" paragraph: The information bottleneck paper of Tishby et al. is not a suitable reference for the general phrase "information theory."
- page 3, last paragraph: you state that v is a probabilistic function of x and noise variable epsilon. Based on later sentences, it seems that g is a deterministic function of x and epsilon.
- Near eq (9), it would have been easier to understand how the method works if you described a few examples of transformation t_i you consider.
- Table 3: inconsistent bolding. L_y=24+w/o Local+MAE, L_y=48+w/o Variety+(MSE, MAE) should be bolded.

Summary Of The Review:
While the paper studies an interesting problem, the reported performance gains are marginal, especially given the complexity of the model. I lean towards rejection.

Correctness: 2: Several of the paper's claims are incorrect or not well-supported.

Technical Novelty And Significance: 2: The contributions are only marginally significant or novel.

Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Flag For Ethics Review: NO.

Recommendation: 5: marginally below the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Add **Public Comment**

[−]

Response to the comments from the reviewer ikmv Part 2

ICLR 2022 Conference Paper947 Authors

30 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=bk4LtiuQyWr)

Comment:

[Comment 4] I am not fully convinced that the proposed paper brings enough benefits to justify its complexity. For example, in Figure 8 and Table 5, the model mostly up-weights the subsequence policy, and the final performance is not very different from only using subsequence.

We have conducted experiments on various datasets to show highly competitive performance with up to 11.4% reduction in MSE on forecasting tasks. The improvement is non-trivial.

With Ablation studies, we verified that 1) the proposed criteria are good to guide the selection of data augmentations. 2) The meta-learner network can successfully select the suitable augmentation based on the proposed criteria. The reviewer points out a phenomenon that ``final performance is not very different from only using subsequence''. The reason is that subsequence augmentation is the most powerful augmentation method for the electricity dataset and achieved the best performance compared to other augmentations. However, subsequence is not always the best for all datasets. That's why we need to adaptively select augmentations that are suitable to the dataset. For example, in the CricketX dataset, the performances of applying each basic augmentation are shown in the following. The results show that Window Slice and Time War are more useful than subsequence. Our InfoTS_s and InfoTS improve over Subsequence by 10.9% and 10.1%, respectively.

InfoTS_s	InfoTS	Cutout	Jittering	Scaling	Time Warp	Window Slice	Window Warp	Subsequence
0.780	0.774	0.679	0.686	0.688	0.739	0.720	0.701	0.703

Without ground truth and other human knowledge, how to adaptively select suitable augmentations is an unsolved problem. That is the main focus of this paper. We try to answer what criteria are and how to find good augmentations based on the proposed criteria.

We thanks for the reviewer's comment and will revise and fix minor comments in the next version.

Add **Public Comment**

[−]

Response to the comments from the reviewer ikmv

ICLR 2022 Conference Paper947 Authors

30 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=e0AKIJ16w1W)

Comment:

[Comment 1] Why is it reasonable to drop H(v)?

H(v) measures the amount of information of (augmented) instances without considering the original instances. This is reasonable for data generation but not for data augmentation. In the literature, all data augmentation methods create new instances based on the original one instead of from scratch. Maximizing H(v) could be achieved by using random noise as augmented instances that have no relationship to the original instances. This kind of augmentation is meaningless for contrastive learning. That's why we dismiss H(v) and focus on -MI(v;x).

[Comment 2] Properties 1 and 2 are fairly trivial statements, which I don't think need proof.

These properties are intuitively correct in case that some readers, such as reviewer 7xBr, have concerns about their correctness. We added the detailed proof in the appendix to make the paper more solid.

[Comment 3]

In this paper, we aim to design a network that can adaptively choose suitable augmentations based on the datasets for downstream contrastive learning. In short, we learn suitable augmentations for contrastive learning. This method is adapting or generalizing to new datasets (environments). That is why we name the network as a meta-learner network. Similar usages are also found in the literature.

Luo, Xu, et al. "Boosting Few-Shot Classification with View-Learnable Contrastive Learning." 2021 ICME, 2021.

[Comment 4] All experiments: seeing error bars would be helpful because the differences between methods are fairly small in some cases.

We compare the proposed method with both time series forecasting and classification. For time series classification, we compare 128 UCR datasets and 30 UEA datasets. It is non-trivial to improve the averaged classification accuracy by 2.8% over the best baseline. We believe that the number of datasets is large enough to verify the consistent superior of the proposed method. That is why error bars are not adopted in the literature. Instead, they report averaged rank to show the averaged performances. Besides, the average rank achieved by the proposed method is 2.133 compared to 2.933 achieved by the best baseline. Thanks for the comments. We will consider updating the manuscript by adding error bars to make it more solid.

[1] Franceschi, Jean-Yves, Aymeric Dieuleveut, and Martin Jaggi. "Unsupervised scalable representation learning for multivariate time series." NeurIPS(2019). [2] Cao, Defu, et al. "Spectral temporal graph neural network for multivariate time-series forecasting." NeurIPS (2021).

[Comment 3] Table 3: these results are so close that I think many of them are not statistically significant. In particular, removing components from the meta objective seems to have minimal effect on final performance.

The main reason is that the basic augmentations are adopted from a previous paper [14], which have been manually tuned to generate relatively good performances for contrastive learning. For these reasons, significant improvements on different variants of our methods are impractical and unreasonable. In fact, in the initial version, we have conducted some experiments on classification which adopts ACC as the metric, using both good and bad basic augmentations. As shown in Figure 3, augmentations with high variety and high fidelity can achieve good accuracy performance with significant improvement. To further take the reviewers' comments into consideration, we have conducted more experiments in Appendix D.3.2 with more low-quality augmentations. The gaps between different variants are more obvious, demonstrating that both variety and fidelity make contributions to the superior performances of InfoTS.

Add **Public Comment**

[−] Summary of authors responses. Please let us know if any questions or additional comments.

ICLR 2022 Conference Paper947 Authors

23 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=Ecuc-ZknNPU)

Comment:

We sincerely appreciate all reviewers for their hard work and helpful comments. We would like to address all reviewers’ main concerns in the corresponding responses. We also updated our manuscript according to the comments. The change we made majorly includes (highlighted in blue in the revised manuscript):

1. We have conducted more experiments in Appendix D.3.2 to comprehensively show the effectiveness of our InfoTS on adaptively selecting optimal augmentations for time series forecasting to further address the reviewers' concerns.
2. We have re-organized related works and added another subsection to introduce these baselines along with their relations to the proposed method in Section 3.
3. We have updated Figure 8 (Figure 7 in the initial version) with updated important weights.

4. We have added more details on selecting positive and negative samples for local contrastive loss in Appendix B.2.
5. We have added more details on the experimental settings in Appendix C.1.

Add **Public Comment**

[−]
Official Review of Paper947 by Reviewer 61sZ

ICLR 2022 Conference Paper947 Reviewer 61sZ

05 Nov 2021 ICLR 2022 Conference Paper947 Official Review Readers: 🌐 Everyone Show Revisions (/revisions?id=7hlDg0gwNM)

Summary Of The Paper:
This is a very good and solid paper that has both empirical elements and theoretical basis. As mentioned in the paper data augmentation in time series is a notoriously difficult problem for the simple reason that it can distort the time series completely - which in contrast to images - humans cannot verify this as it can happen with images. So the contribution of this paper is a new data augmentation approach based on information theory, a meta learning approach and an approach to select optimal data augmentation for contrastive learning.

Main Review:
The paper proposes a new data augmentation approach based on information theory (mutual information) and meta learning to identify optimal data augmentation approach. The rationale behind the method has been extensively described both in the paper and the supplementary material including proofs and additional ablation studies. The paper is experimentally very strong with a comprehensive experiment section and results across various ablation studies. I do not see any weaknesses per se or limitations. Perhaps I missed it somewhere but is there any information on run times ans complexity?

Summary Of The Review:
This is a very good submission that touches upon a kind of neglected and less fancy area of data augmentation for time series - authors propose a very neat solution that improves upon the state of the art and demonstrates good performance across datasets and various ablation studies. I do not think that being too picky on things that do not matter for the sake of finding some minor negative things that do not really affect the scientific merit of the paper. Having said that, the although am working in contrastive learning and self supervised learning, time series data augmentation is not my cup of tea.

Correctness: 4: All of the claims and statements are well-supported and correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Flag For Ethics Review: NO.

Recommendation: 10: strong accept, should be highlighted at the conference

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Add **Public Comment**

[−]
Response to the comments from the reviewer 61sZ

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=t37fo1GDK1p)

Comment:
Thank you so much for the constructive feedback. The proposed meta-network is efficient with the number of parameters linear to the number of candidate augmentation methods. The overall training time is similar to the basic model without Meta-network.

Add **Public Comment**

[−] Official Review of Paper947 by Reviewer 7xBr

ICLR 2022 Conference Paper947 Reviewer 7xBr

04 Nov 2021 (modified: 12 Nov 2021) ICLR 2022 Conference Paper947 Official Review Readers: Everyone Show Revisions (/revisions?id=wdModhMBe8M)

Summary Of The Paper:
This paper describes an information-aware approach to representation learning for time series. The formulation focuses on how to obtain effective data augmentations and addresses the underlying problem from information-theoretic viewpoints, leading to the two optimization criteria, namely, high fidelity and high variety. The experimental results on several time series datasets for forecasting and classification show improvements over the methods in comparison. Detailed comments are listed below.

Main Review:
Detailed comments are listed below.

1. About high fidelity: This criterion requires label information, where in the approach, both ground-truth labels and one-hot encoding pseudo labels have been exploited. Property 1 and Property 2 can be readily proved with the strong assumption of one-hot encoding pseudo labels. That is, the total number of classes is equal to the number of time series samples. While the assumption is hard to implement for practical use, the proposed relaxation reduces the number of (pseudo) classes to the batch size, and consequently weakly supports the validity of Property 1 and Property 2. What are the effects of this relaxation to the proposed approach, or more specifically, to the concept of high fidelity?
2. About high variety: The technical correctness of this part needs to be justified. From (3), seeking an augmentation transform of high variety can be achieved by minimizing $\mathrm{MI}(v; x)$, the mutual information between the augmented instance v and the input x. The subsequent derivation seems to imply that minimizing the negative InfoNCE (as in (5)) can lead to minimize the above mutual information. This is questionable in that minimizing the negative InfoNCE is simply minimizing a lower bound of $\mathrm{MI}(v; x)$; however, it does not ensure that $\mathrm{MI}(v; x)$ will also be minimized.
3. About the meta-contrastive learning in Section 2.3: While the local-wise contrastive loss in (7) follows Tonekaboni et al. (2021), the positive neighborhood and the negative neighborhood should be described in more detail. Also, the subscript i in $\mathcal{N}_i$ is confusing and undefined. The fusion scheme and (9) learn to yield a combined transform and appear to be heuristic. The authors are expected to justify the connection of the proposed approach with meta learning. What are the justifications of naming the architecture as a meta-learner network?
4. The experimental results are not convincing. From the ablation study in Table 3, the results suggest that using the sophisticated fusion scheme (Section 2.3.2) or random augmentations perform almost equally well, while using either “Fidelity” or “Variety” alone does not significantly degrade the MSE outcome at all. In Table 2, as the model InfoTS_s indeed uses label information in training (to select suitable augmentations), the discussion (in the last paragraph of page 8) on outperforming other baselines may not be fair.

Summary Of The Review:
Main concerns about the paper are its technical correctness and the experimental results, including the ablation analysis. This work seems to be not ready for publication yet.

Correctness: 2: Several of the paper’s claims are incorrect or not well-supported.

Technical Novelty And Significance: 2: The contributions are only marginally significant or novel.

Empirical Novelty And Significance: 1: The contributions are neither significant nor novel.

Flag For Ethics Review: NO.

Details Of Ethics Concerns:
None

Recommendation: 3: reject, not good enough

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Add **Public Comment**

[−] Response to the comments from the reviewer 7xBr (Part 2)

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=P3rBcTZLP5D)

Comment:

[Comment 3]. Selection of positive and negative neighborhoods, notation of $\mathcal{N}_i$, and justifications of naming the architecture as a meta-learner network.}

Due to the space limitation, we described the selection of positive and negative neighborhoods in Appendix B.2.

[Q.1] Notation of $\mathcal{N}_i$.

We assume that the reviewer was mentioning the $\bar{\mathcal{N}_i}$ in Eq.(7). We use the $\bar{\mathcal{N}_i}$ as a whole to represent the Non-neighboring samples of a subsequence, which are considered as negatives. Note that the subscript i in $\bar{\mathcal{N}_i}$ is not used in other places. However, we found that replacing the subscript i with s will be more clear, i.e., using $\bar{\mathcal{N}_s}$ to denote the negative samples of subsequence s. We have revised it in our paper. Thanks a lot for the comment.

[Q. 2] Justifications of naming.

In the literature, meta-learning is also known as learning to learn that the model *"is adapting or generalizing to new tasks and new environments that have never been encountered....}"*[9]. We believe that generalizing to new environments is a crucial aspect of meta-learning, which is also well recognized in other meta-learning tutorials such as [10]. In this paper, we aim to design a network that can adaptively choose suitable augmentations based on the datasets for downstream contrastive learning. In short, we learn suitable augmentations for contrastive learning. This method is adapting or generalizing to new datasets (environments). That is why we name the network as a meta-learner network. Similar usages are also found in the literature [11][12].

[9] Lilian Weng, Meta-Learning: Learning to Learn Fast, (OpenAI) https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html (https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html)

[10] Chelsea Finn, Learning to Learn, https://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/ (https://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/)

[11] Lee, Eugene, Evan Chen, and Chen-Yi Lee. "Meta-rppg: Remote heart rate estimation using a transductive meta-learner." ECCV, 2020.

[12] Luo, Xu, et al. "Boosting Few-Shot Classification with View-Learnable Contrastive Learning." 2021 ICME, 2021.

[Comment 4]. The experimental results are not convincing.

[Q.1] Performance in Table 3.

The reviewer points out a phenomenon that *"random augmentations perform almost equally well and "Fidelity'' or "Variety'' alone does not significantly degrade the MSE outcome.''*. First, different from ACC in classification, MSE and MAE are metrics for regression evaluation. As pointed out in previous papers, improvements on these metrics are much harder than ACC (Page 6) [13]. The second reason is that the basic augmentations are adopted from a previous paper [14], which have been manually tuned to generate relatively good performances for contrastive learning. For these reasons, significant improvements on different variants of our methods are impractical and unreasonable. In fact, in the initial version, we have conducted some experiments on classification which adopts ACC as the metric, using both good and bad basic augmentations. As shown in Figure 3, augmentations with high variety and high fidelity can achieve good accuracy performance with significant improvement. To further take the reviewers' comments into consideration, we have conducted more experiments in Appendix D.3.2 with more low-quality augmentations. As shown In Table 8, the performance of ``Random'' is unsatisfactory by using augmentations with lower quality. For example, "Random'' gets 0.384 averaged MAE, while our InfoTS achieves 0.370 average MAE achieved. We also report other variants to further empirically demonstrate that both variety and fidelity make contributions to the superior performances of InfoTS.

[Q.2] Performance in Table 2.

The reviewer points out that *" as the model InfoTS$_s$ indeed uses label information in training (to select suitable augmentations), the discussion (in the last paragraph of page 8) on outperforming other baselines may not be fair."* In the initial version, we have clearly stated the setting of InfoTS_s and fully unsupervised InfoTS. In the last paragraph of Page 8, we also analyzed the results of both InfoTS_s and InfoTS in detail. In the unsupervised setting, InfoTS still outperforms baselines significantly.

[13] Fan, Wenqi, et al. "Graph neural networks for social recommendation." The World Wide Web Conference. 2019. [14] Fan, Haoyi, Fengbin Zhang, and Yue Gao. "Self-Supervised Time Series Representation Learning by Inter-Intra Relational Reasoning." arXiv preprint arXiv:2011.13548 (2020).

Add **Public Comment**

[−] **Weak Connection to Meta Leaning and Unsupportive Ablation Study**

ICLR 2022 Conference Paper947 Reviewer 7xBr

23 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=M3g-6ZE0Zwn)

Comment:

3. About the naming of meta-learner network, I do not think the proposed approach has a strong connection to the meta-learning paradigm. The formulation stated from (9) to (10) is the description of an ensemble scheme to adaptively form an ensemble augmented instance. Particularly, the proposed scheme uses Bernoulli sampling to generate a transformation, and then considers an averaged ensemble over multiple such transformations to yield an augmented instance. When dealing with a new dataset, the whole network model needs to be trained either with the previous model parameters as initialization or from scratch all over again. On the other hand, the formulation in the mentioned paper [11] provides **fast adaptation** to the distribution change, exhibiting the typical feature of a gradient-based meta-learning approach.
4. The focal point of this research effort is to propose a way of how to generate augmented instances for time-series representation learning. However, from the provided ablation results in Table3 and the author response, I still feel that my previous comments on the experimental results are justifiable. After all, unsupervised representation learning is typically time consuming. With such complexity, it is reasonable to expect more significant performance gains from the proposed sophisticated ensemble scheme to generate augmented instances at each training iteration, when comparing with simply using random augmentations.

[11] Lee, Eugene, Evan Chen, and Chen-Yi Lee. "Meta-rppg: Remote heart rate estimation using a transductive meta-learner." ECCV, 2020.

Add **Public Comment**

[−] **Clarification of usage of Meta learning and ablation studies.**

ICLR 2022 Conference Paper947 Authors

24 Nov 2021 (modified: 24 Nov 2021) ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=6wR_Y2xGlB_)

Comment:

1. Connection to Meta Leaning

In this paper, we focus on learning to augment for contrastive learning and propose a meta-network to adaptively select suitable augmentations for contrastive learning. We believe that this usage also follows the paradigm of learning to learn. For the "fast adaption" concern, a trained meta-network also has the potential to be used in new datasets with fine-tuning. We leave this extension in our future work. Thanks for the insightful comments.

2. Unsupportive Ablation Study

In the previous review, the reviewer pointed out a phenomenon that the improvement is not significant compared to "Random" and our two variants. We analyzed the phenomena and added more experiments in Appendix 3.2 to further demonstrate the effectiveness of the proposed method in practice. We thank the reviewer for the insightful comments, such that we can make our paper more well-qualified.

The reviewer still believes that our ablation studies are unsupportive. To take the reviewer's concern, we summarized our major and minor contributions as follows.

The two main contributions of our paper are 1) proposing criteria of data augmentations for time series based on information theory that good augmentations should have high variety and fidelity 2) proposing a meta-network to adaptively select augmentations for each dataset based on the proposed criteria.

For the first contribution, we have verified with Figure 3 in Section 3 and Figure 9 in Appendix 3.2. For the second contribution, we have verified with Table 3 and Table 8 that we outperform over "Random," which is used in previous works. The margins become more significant if there is no prior knowledge on the selection of augmentations (Appendix 3.2). This phenomenon shows that InfoTS is more practical because the quality of augmentations on time series data is not easy to check by humans virtually. Besides, The comparison of InfoTS to its variants also demonstrated that both fidelity and variety are important for the selection of augmentations. Another minor contribution in this paper is the contrastive loss. We have also verified its effectiveness with ablation studies.

Thus, we believe that we have adequately verified our claims with extensive experiments. As we analyzed in the experimental part, experimental results also support our claims and we respectfully disagree with the reviewer's comment on "unsupportive ablation study".

Add **Public Comment**

[−] **Response to the comments from the reviewer 7xBr**

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=dPg_YHszGl)

Comment:

Thank you so much for the constructive feedback. We sincerely appreciate your valuable suggestions and questions. The followings are our responses.

[Comment 1]. The motivation of relaxation and its effects on the concept of high fidelity.

As shown in Figure 2(b), in the unsupervised setting, keeping high fidelity requires that ``the generated instances are constrained to the region around the raw input. Such that they are still distinguishable from other instances.'' This can be achieved by using one-hot encoding as the pseudo label, which considers ALL other instances. However, such implementation is inefficient with poor scalability for practical use. A routinely used method is negative sampling, which conducts random sampling for negatives. In our setting, we adopt other instances in the same batch as the negative samples. This kind of approximation is broadly used in contrastive learning methods like SimCLR[1], which achieves good performances in various tasks.

Theoretically, we consider the output of the classifier as a distribution over B classes, where $B > 1$ is the batch size. Then, an (ideal) classifier that trained with mini-batch still maintain the property that " $f_\theta(x) = f_\theta(v)$ if and only if v is an augmentation of instance x". Such property guarantees the one-to-many map between variable x and v, which also supports our claims on Property 1 (Preserving Fidelity) and Property 2 (Adding New Information).

[Comment 2]. The technical correctness of high variety.

We recognized that the motivation of using InfoNCE as an approximation of MI was originated from the property that InfoNCE is a lower bound of MI. However, in recent works [2][3][4][5][6], InfoNCE has also been extended to a general estimation method for MI, whether it is in the forms of minimizing or maximizing, due to its practical good performances. In [7], the authors have summarized this kind of usages. *"Minimizing statistical dependency in representations is a common goal in disentangled representation learning. Prior work has focused on two approaches that both minimize lower bounds: (1) using adversarial learning (Kim & Mnih,2018; Hjelm et al., 2018), or (2) using minibatch approximations where again a lower bound is minimized (Chen et al.,2018)}"*. A more recent example is InfoMin [2] in NeurIPS 2020, they say that *"We use I_{NCE} as a neural proxy for I. (Page 5)}''*. Based on this approximation, they model the (augmented) view learning problem as an adversarial min-max problem.

The underlying theory of the usage has been analyzed in previous works [7][8]. In summary, a small modification of InfoNCE (leaving the positive one out in the denominator) can generate an upper bound estimation of MI. However, by doing so, it *``suffers from numerical instability in MI optimization and fails during training. (Page 7)}''* [8].

Thus, in this paper, we follow previous works and choose InfoNCE as an estimation of MI. Besides, As we have discussed in the initial version, our framework is flexible in the selection of MI estimators. Other methods can also be adopted here. Thanks a lot for your comments. We have revised our paper accordingly to make it more precise.

[1] Chen, Ting, et al. ``A simple framework for contrastive learning of visual representations.'' ICML, 2020.

[2] Tian, Yonglong, et al. "What makes for good views for contrastive learning?." NeurIPS (2020).

[3] Kim, Hyunjik, and Andriy Mnih. "Disentangling by factorising."ICML, 2018.

[4] Hjelm, R. Devon, et al. "Learning deep representations by mutual information estimation and maximization." ICLR 2019.

[5] Chen, Ricky TQ, et al. "Isolating sources of disentanglement in variational autoencoders." NeurIPS (2018).

[6] Suresh, Susheel, et al. "Adversarial Graph Augmentation to Improve Graph Contrastive Learning." NeurIPS (2021).

[7] Poole, Ben, et al. "On variational bounds of mutual information." ICML, 2019.

[8] Cheng, Pengyu, et al. "Club: A contrastive log-ratio upper bound of mutual information." ICML, 2020.

Add **Public Comment**

[−] Technical Correctness Needs Further Justifications.

ICLR 2022 Conference Paper947 Reviewer 7xBr

23 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: Everyone Show Revisions (/revisions?id=XgV0r77qqmR)

Comment:

1. About high fidelity: In the response, it is stated that when considering the output of the classifier as a distribution over B classes, where $B > 1$ is the batch size, Property 1 and Property 2 are still valid. This statement is questionable. The proof of Property 1 relies on the validity of $p(v \mid x, y) = p(v \mid x)$. However, this holds only for the case that the total number of classes is equal to the number of time series samples. It would be more insightful if the authors can discuss the effects of the weak relaxation to the theoretical (but impractical) assumption due to drastically reducing the total number of classes.
2. About high variety: My main concern is that minimizing mutual information $\mathrm{MI}(v; x)$ cannot be guaranteed or enforced by minimizing its lower bound (in this case, the negative InfoNCE). In the response, the authors quote from [7] the following to support their approach, namely, “*Minimizing statistical dependency in representations is a common goal in disentangled representation learning. Prior work has focused on two approaches that both minimize lower bounds: (1) using adversarial learning (Kim & Mnih,2018; Hjelm et al., 2018), or (2) using minibatch approximations where **again** a lower bound is minimized (Chen et al.,2018).*”

The reference to [7] raises more questions about the correctness of their approach. If one reads the whole paragraph of the section entitled “Upper bounding total correlation” of [7], the discussions there seem to imply the inappropriateness of minimizing a lower bound in previous approaches. The next sentence (unquoted) in the paragraph **right after** those quoted in the author response instead says that “**To measure and minimize statistical dependency, we would like an upper bound, not a lower bound**.”

I would appreciate if the authors can directly provide theoretical arguments to justify **why minimizing** $\mathrm{MI}(v;x)$ **can be achieved by minimizing its lower bound**.

[7] Poole, Ben, et al. "On variational bounds of mutual information." ICML, 2019.

Add **Public Comment**

[−]

Clarification of on the reviewer's concerns on the technical correctness

ICLR 2022 Conference Paper947 Authors

24 Nov 2021 (modified: 24 Nov 2021) ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=HbZsv1F0_tt)

Comment:

Thank you so much for the quick feedback. We sincerely appreciate your valuable comments. The followings are our responses.

[Comment 1] The the validity of $p(v|x,y) = p(v|x)$.}

As we have explained in the appendix of both initial and revised versions, *"In the unsupervised setting where the ground-truth label y is unknown, we assume that the augmentation v is a (probabilistic) function of x only. The only qualifier means $p(v|x,y) = p(v|x)$."* Here y is the ground-truth label rather than the pseudo label.

The validity of $p(v|x,y) = p(v|x)$ is from assumption, and it is not related to mini-batch approximation. This assumption is easy to understand. Using the subsequence for an example, we generate an augmented instance v from an original instance x. This augmentation process is independent of the ground-truth label of x in the unsupervised setting. That's why *"we assume that the augmentation v is a (probabilistic) function of x only."*

[Comment 2]] About the high variety

We thank the reviewer for the insightful comment on the usage of InfoNCE as an estimator of MI. As we have responded early, using InfoNCE as an estimation method for MI, whether it is in the forms of minimizing or maximizing, is wide-recognized in the literature. Some recent works, such as [2][6] (two recent NeurIPS papers), also adopt this kind of usage for its practical effectiveness, even after [7] was accepted. Even though, theoretical guarantee that "minimizing MI can be achieved by minimizing its lower bound (InfoNCE)" has not been rigorously analyzed in the literature. We agree that it is very important to theoretically analyze the relationship between minimizing InfoNCE and minimizing MI because of its wide applications in various tasks. However, this is beyond the scope of this paper. Note that, in our initial version, we introduced the usage without any incorrect claims. *"we adopt a mutual information neural estimator, InfoNCE, to approximately compute the mutual information."* To make it more clear, we have already revised our paper to clearly state that we choose InfoNCE because of its practical effectiveness. Other approximations with rigorous theoretical guarantees, such as leave-one-out, can also be adopted. Thanks for the reviewer's comment.

[2] Tian, Yonglong, et al. "What makes for good views for contrastive learning?." NeurIPS (2020).

[6] Suresh, Susheel, et al. "Adversarial Graph Augmentation to Improve Graph Contrastive Learning." NeurIPS (2021).

[7] Poole, Ben, et al. "On variational bounds of mutual information." ICML, 2019.

Add **Public Comment**

[–] **Official Comment by**

[Deleted]

ICLR 2022 Conference Paper947 Official Comment Readers: Everyone Show Revisions (/revisions?id=im5az2T2wl)

[–] **Official Review of Paper947 by Reviewer ZzdQ**

ICLR 2022 Conference Paper947 Reviewer ZzdQ

26 Oct 2021 (modified: 26 Oct 2021) ICLR 2022 Conference Paper947 Official Review Readers: Everyone Show Revisions (/revisions?id=_6fca2LP2sO)

Summary Of The Paper:

The authors propose a contrastive learning framework for time series data, where data augmentations are adaptively being selected, given a fidelity and variety criterion. Additionally to these two criteria, a contrastive learning objective is applied both on local and global level. The model is tested on a time-series forecasting and classification task on multiple datasets.

Main Review:

Strengths: The authors propose an interesting solution to the challenge of seeking data augmentations on time-series data and provide a bunch of experiments on various datasets and with different benchmark models. The paper is well-written.

Weaknesses: Some of the analyses and ablations could be more complete, in order to truly investigate whether the conclusions are correct.

More specific feedback:

- It is unclear to me what the definition of the pseudo-labels is. Does it mean, if we have, e.g., 10k data points, that the pseudo label is a one-hot encoding with 10k elements? So each data points has its own 'class' label? A proper definition is lacking.
- p4: The authors mention that direct optimization of Eq. 2 is inefficient and unscalable. But I don't really understand why, as it is in fact just a classification model like any other supervised classifier. So why is it impossible to use the label for each datapoint? And then second, how is it justified that the label of the full batch y_B does contain information about all data windows in the batch? What is it's definition? This addition to the model feels a bit strange, as it is also not further mentioned in the rest of the paper.
- eq. 10: Are the p_i probabilities trainable parameters? And how do they relate to the importance scores mentioned in appendix D.3? Are the importance scores just unnormalized probabilities? How is the temperature set during training?
- 3.1: This paragraph is a summation of a lot of models, but it does actually not explain the models that are finally used as benchmark models. If the reader is unfamiliar with any of the benchmark models, the paper does not provide any information to the reader about their difference compared to the proposed model. So I would extend this section with a (semi-technical) explanation of the benchmark models and how they differ in their main technique. For example, which of the proposed benchmarks is actually also a contrastive learning model, and which of the benchmarks is supervised? This information is useful to better interpret the provided results.
- Contrastive Predictive Coding (van Oord et al., 2018) is another well-known contrastive learning method that is often used for time-series, but the authors did not mention it anywhere in the paper. Besides mentioning it in related work, it would be a very useful benchmark as well, possibly combined with the local and global loss as the authors proposed.
- p7, last line: The authors mention that TS2Vec achies 2nd best perofrmance because it adopts subsequence by random cropping. How do the authors know this is the reason? No causal conclusions can be drawn from the presented Tabels.
- Same remark for p8, where the authors argue that InfoTS can adaptively select the most suitable augmentations. At this point of reading, there is no proof of that yet. Later in the ablation studies the authors do however check for this, but the conclusion from Table 3 is mainly that all factors contributed (at least a bit), where the Fidelity and Variety criterion seemed to have contributed the least. So it's not super fair to state that the model's superior performance is (only) thanks to adaptive augmentation selection. In fact, from Fig. 7 it even

seems that the model just learns to select one best augmentation, rather than a best combination of all of them.

- Fig. 3: Can the authors also provide the plot for the ablation models that are not trained with the variety and fidelity criteria? Now it's unsure whether these criteria contributed to this positive relation, or whether this relation already exists by default in such models. Also, how does this relation look like in the forecasting task?

Minor things/ typos:

- end p3: "parameterized" instead of paramterized
- I would refer to Fig. 2 already earlier, for example already both in the High fidelity and high variety sections. It helps the understanding of the reader.
- p4: The authors mention that the nr of labels is equal to the number of instances in dataset X in the unsupervised case, but this is actually for supervised training right? In the unsupervised case there are no labels. Or do the authors refer to the pseudolabels then? But even then, in the supervised case this remark also holds.
- p5: Is the batch X_b the same as the mini-batch used during training and mentioned later in the paper? Or are batch and mini-batch two different entities in this work?
- p5: "non-neighboring samples", is a sample here a subsequence, just like s?
- eq. 10: The work from Jang et al., ("Categorical reparameterization with gumbel-softmax") concurrently invented the concrete/Gumbel-softmax distribution as the work of Madisson et al., so these works are typically cited together.
- LSTnet is compared in table 1, but not cited in section 4.1, and StemGNN is cited but not added to the table.
- 4.4: "The advantage of adaptive selection": So if I understand correctly with adaptive selection the authors refer to the concrete sample from a trained categorical over possible augmentations?
- Appendix C.1: What's the difference between the cutout and the subsequence augmentation?
- Appendix C.1: If time warping is applied, the number of samples per window is different right? So how do the authors deal with this? Is resampling applied maybe?
- Appendix D.1: To which experiment does figure 5 relate? And are all runs run with the same randomized seed?
- Appendix D.3: The MAE of Subsequence for L_y = 168 and 336 in Table 5 are missing a leading zero.

Summary Of The Review:
The authors propose an interesting solution to the challenge of seeking data augmentations on time-series data. Downstream task performance seems promising compared to benchmark models. In general the paper is well-written, but some clarifying questions remain (see my earlier remarks). Also, some of the (ablation) analyses can be made more complete.

Correctness: 3: Some of the paper's claims have minor issues. A few statements are not well-supported, or require small changes to be made correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: 4: The contributions are significant, and do not exist in prior works.

Flag For Ethics Review: NO.

Recommendation: 6: marginally above the acceptance threshold

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Add **Public Comment**

[−]

Response to the comments from the reviewer ZzdQ (Part 3)

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=R3kX_NOrV7W)

Comment:

[Comment 9] P5

####[Q. 1] *p5: Is the batch X_b the same as the mini-batch used during training and mentioned later in the paper? Or are batch and mini-batch two different entities in this work*

Thanks for pointing it out. Mini-batch and batch are the same entities in this work. A (mini-batch) of instance is denoted by $\mathbb{X}_b$.

[Q,2] p5: ``non-neighboring samples'', is a sample here a subsequence, just like s?

Yes. We have added more details about how to select positive and negative samples in Appendix B.2.

[Comment 10] Experiments

[Q.1] *LSTnet is compared in table 1, but not cited in section 4.1, and StemGNN is cited but not added to the table.*

Thanks for the comments. I think in the initial version, we have already introduced these methods with citations. StemGNN is not included in Table 1 (Univariate time series forecasting) because *"StemGNN is for multivariate only"*. We have re-organized the order of compared methods in Section 4.1 to make it consistent with Table 1.

[Q.1] *4.4: The advantage of adaptive selection: So if I understand correctly with adaptive selection the authors refer to the concrete sample from a trained categorical over possible augmentations?*

Yes, our meta-network learns a weight for each candidate augmentation, which is considered as adaptive selection.

[Comment 11] Appendix

[Q.1] *Appendix C.1: What's the difference between the cutout and the subsequence augmentation?*

In most cases, the cutout operation generates two subsequences connected by an all-zero sequence. Subsequence augmentation generates a subsequence padding with zeros at the beginning and the end. In most cases, we find that subsequence is more useful than the cutout.

[Q.2] *Appendix C.1: If time warping is applied, the number of samples per window is different right? So how do the authors deal with this? Is resampling applied maybe?*

As we have introduced in the initial version, we adopt the tool from https://tsaug.readthedocs.io (https://tsaug.readthedocs.io). With this function, we may generate augmented sequences with different lengths. We further apply over-sampling or sampling to make sure the length is the same as the original one. We have added more details here. Thanks for pointing it out.

[Q.3] *To which experiment does figure 5 relate? And are all runs run with the same randomized seed*

Thanks for the comments. This part is following Section 4 and thus the setting is consistent with other parts related to the Electricity dataset. All runs were run with the same randomized seed.

We thank you for other comments from Review ZzdQ and have modified the submission accordingly.

Add **Public Comment**

[−]

Response to the comments from the reviewer ZzdQ (Part 2)

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=0XkgJESdAyS)

Comment:

[Comment 6] Remark on P8.

[Q.1] Performance in Table 3. *"the Fidelity and Variety criterion seemed to have contributed the least.''*.

The reason is that the basic augmentations are adopted from a previous paper [14], which have been manually tuned to generate relatively good performances for contrastive learning. For these reasons, significant improvements on different variants of our methods are impractical and unreasonable. In the initial version, we have conducted some experiments on classification, using both good and bad basic augmentations. As shown in Figure 3, augmentations with high variety and high fidelity can also achieve good accuracy performance with significant improvement. To further take the reviewers' comments into consideration, we have conducted more experiments to show the effectiveness of each component with both good and bad basic augmentations in Appendix D. As is shown In Table 8, comparing to "Random'', which is not trained with variety and fidelity objectives, our InfoTS significantly increases the performances. For example, "Random'' gets 0.384 averaged MAE, while our InfoTS achieves 0.370 average MAE achieved. We also report other variants to further empirically demonstrate that both variety and fidelity make contributions to the superior performances of InfoTS.

[Q.2] Regarding Figure 7. *it even seems that the model just learns to select one best augmentation, rather than a best combination of all of them".

Figure 7 shows the weight updating process of the meta-network on the Electricity dataset. In the initial version, we didn't normalize weights. We have updated the figure with normalized weights in the revised version. The most important conclusion from the figure is that the optimal augmentation selected by our meta-network is consistent with the one selected with testing accuracy performance.

As explained in Section 2.3.2, each candidate augmentation is associated with a weight, and the *"adaptive augmented instance can be achieved by combing candidate ones'' (Page 6)*. In other words, theoretically, the meta-network considers the combination instead of just selecting the best one. In addition, we have empirical observations in Table 5, InfoTS outperforms the variant that uses subsequence only (the last column). This comparison empirically shows that the meta-network learns to consider the combinations, which is better than any (single) candidate augmentation.

[Comment 7] Fig.3.

We first briefly summarize the logic of our paper as its experimental part. The two main contributions of our paper are 1) propose criteria of data augmentations for time series based on information theory that good augmentations should have high variety and fidelity 2) we propose a meta-network to adaptively select augmentations for each dataset based on the proposed criteria. Figure 3 is used to empirically verify the first contribution. This part is independent of the (trained) meta-network. Each dot represents an augmentation with a unique hyper-parameter configuration. The augmentation part is not trained with any criteria. The most important conclusion is that *"in general, accuracy performance is positively related to the proposed criteria in both supervised and unsupervised settings.''* We have revised the analysis to make it more precise. For the second contribution, we empirically verified it with Section 4.4 and appendix D (More experiments have been added in the revised version).

[Q.1] *Can the authors also provide the plot for the ablation models that are not trained with the variety and fidelity criteria"*.

This variant is similar to the variant "Random", which is not trained with any criteria, in Table 3 and Table 8.

[Q.2] *how does this relation look like in the forecasting task ?*.

We have plotted the relations in Figure 9 in the appendix. The results are consistent with the conclusion drawn from classification results that forecasting performance, evaluated with both MSE and MAE, is positively related to the proposed criteria.

Response to Minor things/ typos:

[Comment 8] P4

[Q.1] *The authors mention that the nr of labels is equal to the number of instances in dataset X in the unsupervised case, but this is actually for supervised training right? In the unsupervised case there are no labels. Or do the authors refer to the pseudo labels then? But even then, in the supervised case this remark also holds.*

Here, we refer to the pseudo labels, which means that the label is not ground truth. We adopt one-hot encoding as the (pseudo) label, where the number of labels equals the number of instances in the dataset. It can also be used in the supervised setting. However, it is unnecessary because we have the ground-truth labels in the supervised case, which provides additional information for selecting optimal augmentations.

Add **Public Comment**

[−] Response to the comments from the reviewer ZzdQ

ICLR 2022 Conference Paper947 Authors

22 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: 🌐 Everyone Show Revisions (/revisions?id=etr8_qRhWoj)

Comment:

Thank you so much for the constructive feedback. We sincerely appreciate your valuable suggestions and questions. The following are our responses.

[Comment 1] Pseudo-labels

Thanks for the comments. Your understanding is correct. We have added proper definitions to make the manuscript more self-contained.

[Comment 2] Optimization of Eq. 2 and usage of batch.

[Q.1] why is it impossible to use the label for each data point?

As we explained later in our paper, *"Since the number of labels is equal to the number of instances in dataset X in an unsupervised case, direct optimization of Eq. (2) is inefficient and unscalable. (Page 4).''* For example, if we have 10k data points, meaning that the pseudo label is a one-hot encoding with 10k elements, such that the output dimension of the classifier will be 10k, which is inefficient with poor scalability.

[Q.2] Justification of usage of batch.

As shown in Figure 2(b), in the unsupervised setting, keeping high fidelity requires that *"the generated instances are constrained to the region around the raw input. Such that they are still distinguishable from other instances.''* This can be achieved by using one-hot encoding as the pseudo label, which considers ALL other instances. However, such implementation is inefficient with poor scalability for practical usage. A routinely used method is negative sampling, which conducts random sampling for negatives. In our setting, we adopt other instances in the same batch as the negative samples. This kind of approximation is broadly used in contrastive learning methods like SimCLR[1], which achieves good performances in various tasks.

[1] Chen, Ting, et al. ``A simple framework for contrastive learning of visual representations.'' ICML, 2020.

[Comments 3] Eq. 10

[Q.1] Are the p_i probabilities trainable parameters? And how do they relate to the importance scores mentioned in appendix D.3? Are the importance scores just unnormalized probabilities?

As shown in Algorithm in Appendix B, $\{q_i\}_{i=1}^{|T|}$ are the parameters in meta-network. After normalization with the sigmoid function, for each q_i, we get the *"important weight $p_i \in (0,1)$, inferring the probability of selecting transformation t_i. (Page 6)''*.

[Q.2] How is the temperature set during training?.

We follow the practice in [2] to adopt the strategy by starting the training with a high temperature and annealing to a small value with a guided schedule.

[1] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICML, 2020. [2] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax". arXiv preprint arXiv:1611.01144, 2016.

[Comment 4] Related Work

Thanks a lot for your kind suggestions. In this paper, we focus on learning time series representations, which be used for forecasting and classification. For this reason, we didn't introduce time series baselines in related work in our initial version. Considering that we also compare with many time series forecasting methods in our experiments, We have re-organized related works and added another subsection to introduce these baselines along with their relations to the proposed method.

Thanks for bringing the work Contrastive Predictive Coding to our attention. We have added this work in the revised version.

[Comment 5] Performance of TS2Vec.

[Q. 1] *p7, last line: The authors mention that TS2Vec achieves 2nd best performance because it adopts subsequence by random cropping. How do the authors know this is the reason? No causal conclusions can be drawn from the presented Tables.*

Thanks for the comments. The most important components of TS2Vec are random cropping and contrastive learning. That's why we draw the causal conclusions in the initial version. However, for more rigorous statements, we have revised this part accordingly. Thanks a lot for the comments.

Add **Public Comment**

[−] **Replies to rebuttal**

ICLR 2022 Conference Paper947 Reviewer ZzdQ

24 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: Everyone Show Revisions (/revisions?id=Jm6b0UxGH8)

Comment:

I thank the authors for the extensive rebuttal and the responses to my questions. Many questions were answered, however, there are some remaining points that I will explain below:

- Comment 2: Thanks for your explanation, I think I initially misunderstood the concept of using the batch (only) for defining the labels. I initially understood that all data points within 1 batch were assigned with the same label. Now I understand that you create one-hot pseudo-labels, where the number of classes equals the number of unique classes in that batch. So y_s \in R^|x_b| right?
- Comment 3: Okay got it. Maybe good to add p_i to the y-axis of (the new) Fig. 8. I would also be more precise in the added comment about the temperature annealing scheme. How is 'high value' and 'low value' defined and which schedule do you use? The way it is reported now, the reader is still not able to exactly reproduce the experiments. Of note, something else that I noticed with respect to reproducability and preciseness of rapportation is about the datasets. You mention that you used the Electricity dataset, but in the ICU database, there are multiple electricity-related datasets. Do you mean the UCI Individual Household Electric Power Consumption Dataset? Please be precise such that reader can truly reproduce results.
- Comments 4: I think the related work did indeed improve from the added discussion on the benchmark models. I see you added the work from Oord et al., 2018, as a reference, but an explanation lacking still, while you do explain the other methods in that section.
- Comment 7: Even though the ablation study Table 3 shows that each of the components (local loss, global loss, fidelity and variety criteria) contributed to the final result, I do not understand why the authors (only) stress the importance of the adaptive augmentations, since the errors for the ablations w/o Local and w/o Global are higher than for the w/o Fidelity and w/o Variety ablations. In my opinion, this indicates that most performance gain can be attributed to the addition of the local and global contrastive losses. But to acuratelly test the gains from the contrastive losses vs the gains from the adaptive augmentations, a fair comparison against TS2VEC is needed, where first the losses of TS2VEC will be replaced by the global and local objectives, and in another experiments the augmentations are changed for the adaptive augmentations. In that way, only one variable changes at a time and fair conclusions can be drawn. In the current setup, the difference with Ts2VEC is twofold: both the losses and the data augmentation strategy has changed. Therefore, another much more additional comparison would be the comparison to (plain) Contrastive Predictive Coding (CPC), and CPC that also includes the additional local contrastive loss. I suspect that CPC with this local loss is possibly performing very similar to the currently presented results.

To conclude: The work has improved from the rebuttal iteration, but I still have concerns regarding the claims and conclusions. The authors pose their work as if adaptive augmentation selection is the key to performance boost in time series representation learning, while their local and global loss objectives seem to be very crucial as well. So the reader can be given false expectations about the performance gains of adaptive data augmentation selection.

Add **Public Comment**

[−] **Clarification of on the reviewer's concerns.**

ICLR 2022 Conference Paper947 Authors

30 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: Everyone Show Revisions (/revisions?id=vLwMwITmOF8)

Comment:

Thank you so much for the constructive feedback. The following are our follow-up responses.

[Comment 2]

Yes. Your understanding is correct.

[Comment 3]

Thanks for the suggestion. The 'high value' and 'low value' are set to 2.0 and 0.1, respectively. Temperature is tuned by the following formula.

```
high * np.power(low / high, current_epochs / total_epochs)
```

For the electricity dataset, the full name is ElectricityLoadDiagrams20112014 Data Set. We download it from https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014 (https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014).

[Comment 7] Even though the ablation study Table 3 shows that each of the components (local loss, global loss, fidelity, and variety criteria) contributed to the final result, I do not understand why the authors (only) stress the importance of the adaptive augmentations since the errors for the ablations w/o Local and w/o Global are higher than for the w/o Fidelity and w/o Variety ablations.

The two main contributions of our paper are 1) proposing criteria of data augmentations for time series based on information theory that good augmentations should have high variety and fidelity 2) proposing a meta-network to adaptively select augmentations for each dataset based on the proposed criteria. Local and global contrastive losses are also a contribution in this paper, although we didn't emphasize it in the submission. We agree with the reviewer's comment that both newly designed contrastive loss and adaptive selections make significant contributions to the Superior performances of InfoTS. We follow the reviewer's suggestion to conduct more experiments to show the effectiveness of adaptive selection with TS2vec as the backbone. The comparison between TS2vec and TS2vec+adaptive shows that adaptive selection can also boost the performance of InfoTS significantly, 10.9% on Screentype and 32.23% on ethanollevel. Since we adopt the same encoder as TS2vec, by replacing local and global contrastive losses in TS2vec+adaptive, we get our method InfoTS, which further improves the performance (8.1% on Screentype and 11.25% on ethanollevel).

Dataset	TS2vec	TS2vec+adaptive	InfoTS
Screentype	0.411	0.456	0.493
ethanollevel	0.484	0.640	0.712

Add **Public Comment**

[−] **Response to authors**

ICLR 2022 Conference Paper947 Reviewer ZzdQ

30 Nov 2021 ICLR 2022 Conference Paper947 Official Comment Readers: Everyone Show Revisions (/revisions?id=8lYsTvo8_Xe)

Comment:

Thanks for the additional information on datasets and settings, this information should also be added in the paper though.

Then with regards to contribution; the authors write they didn't choose to provide the additional local loss as a contribution, I still think it is important to stress to the reader that the main performance gains is thanks to both adaptive augmentation selection and global + local loss. Just choosing to not mention one of the reasons for performance gains does in my opinion still distract the reader from the actual message.

I appreciate the effort of the authors to add additional experiments. This comparison is truly a useful one in general, as only one variable changes at a time. Though, as seen from Table 10, the different classification problems in the 128 UCR dataset perform quite differently when comparing InfoTS to TS2vec. So just reporting only two of these 128 datasets does not tell me that in general the gap between TS2vec and InfoTS is as large as shown in the comment. So this comparison is the most important and should be reported in the major parts of the work.

Add **Public Comment**

View 1 More Reply →

About OpenReview (/about)