Leveraging Time Irreversibility with Order-Contrastive Pre-training

Monica Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag

MIT CSAIL

February 6, 2023

Presented by Matthew Engelhard

Introduction

The authors propose a novel form of self-supervision called **order-contrastive pre-training** (OCP).

Goal: to learn representations of time-series segments sufficient to order any two segments correctly.[1] This encourages the learning of 'time-irreversible' features, i.e. features that only change in one direction over time.

- Similar to self-supervision in other domains (e.g. vision, language), we're motivated by settings/problems in which data is plentiful, but labels are scarce.
- Why do we care about 'time-irreversible' features? They're motivated by downstream tasks for which these are the relevant features – think disease progression, diagnoses, gradual health decline.
- Intuitively, the OCP objective forces the model to identify these features, and they have theoretical results to support this intuition (under strong assumptions, of course).
- The method is **not** appropriate for all clinical prediction tasks, notably ones for which the relevant underlying features or risk factors are transient.

[1] More precisely, to determine whether two segments are correctly versus incorrectly ordered

Helpful Background

This paper is fairly approachable and self-contained, but it would be helpful to have an understanding of general principles of embeddings, self-supervised learning, and contrastive learning.

A few suggestions to get started:

- Here's a nice blog post introducing self-supervised learning.
- For contrastive learning, SimCLR might be a good place to start.

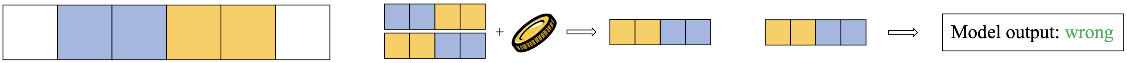
Methods: Setup

Setup: Suppose we have data $\{X_i\}_{i=1}^N$, where each $X_i = (X_i^1, \ldots, X_i^{\tau_i})$ is a time-series with samples $X_i^t \in \mathcal{X}$ whose length τ_i may vary between individuals.

Goal: Learn parameters θ of an encoder g_θ such that the representations $g_\theta(X)$ are (a) sensitive to event ordering, and (b) useful for downstream (i.e. subsequent) tasks.

We structure the learning task as follows:

1. Given trajectory X, sample $W \in \{1, \ldots, \tau - 1\}$ uniformly at random and flip a coin $Y \sim \text{Bern}(0.5)$.
2. If heads ($Y = 1$; correct order), present $(X^W, X^{W+1}, 1)$ as a sample for the contrastive task.
3. If tails ($Y = 0$; incorrect order), present $(X^{W+1}, X^W, 0)$ as a sample for the contrastive task.



1. Sample random pair of *windows* (window1, window2).
2. Flip a coin to decide which way to order the windows (right or wrong).
3. Train the model to detect when windows are given in the correct order.

Figure: Order-contrastive pre-training setup (Agrawal et al.)

Methods: Training Objective

Along with the encoder $g_\theta : \mathbb{R}^{d_\mathcal{X}} \to \mathbb{R}^{d_\mathcal{R}}$, we have classifier $c_\phi : \mathbb{R}^{d_\mathcal{R}} \times \mathbb{R}^{d_\mathcal{R}} \to (0, 1)$. The former will be retained for subsequent tasks, whereas typically the latter will be discarded after pre-training.

The learning objective is structured as follows:

$$\operatorname*{argmin}_{\theta,\phi} \mathbb{E}\left[\mathcal{L}(Y, c_\phi\left(g_\theta(X^{W+1-Y}), g_\theta(X^{W+Y})\right))\right], \tag{1}$$

where $\mathcal{L}$ is the standard cross-entropy (i.e. logistic) loss. Note that the approach is easily extended to use window sizes larger than 1.

The encoder and classifier will later be instantiated as neural networks with appropriate activation functions in their final layers; the objective is then optimized via SGD.

Brief overview of theoretical results

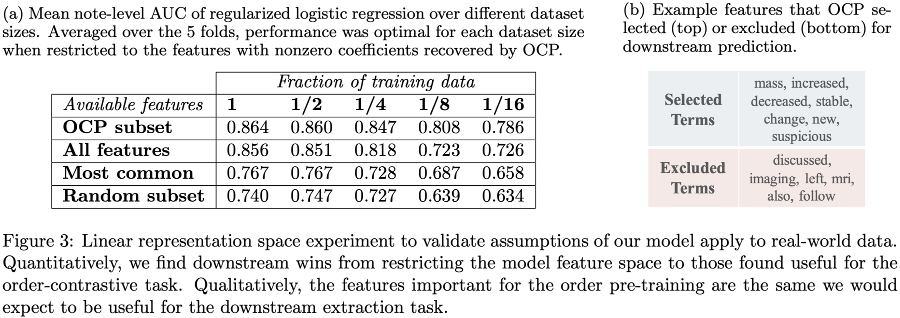
Consider the following setting:

- Our domain consists of d binary features (i.e. $\mathcal{X} = \{0,1\}^d$).
- A subset S of the features cannot decrease (i.e. cannot disappear once present): $\forall i \in S$ and $\forall t$, $\mathbb{P}(X_i^{t+1} < X_i^t) = 0$. They call such features 'time-irreversible'.
- Other features are either (a) noisy indicators of features in S, or (b) background, 'time-reversible' features that are independent of features in S. See paper for precise definitions.
- Note that they define 'time-reversible', i.e. $\mathbb{P}(X_i^t = v, X_i^{t+1} = v') = \mathbb{P}(X_i^t = v', X_i^{t+1} = v)$, to be stronger than simply NOT time-irreversible (i.e. potentially decreasing).
- **The downstream task of interest depends only on the time-irreversible features**. Thus, the goal of pre-training is to identify them.

In this setting, they prove that excess risk (over the optimal classifier) for the downstream task is lower when using order-contrastive pre-training. In intuitive terms, OCP is a good idea if you expect downstream prediction tasks to depend primarily on attributes that do not disappear once present (e.g. diagnoses, disease status, risk factors).

Experimental Results, Part 1

- Task: predict cancer progression from patient notes via (1) OCP-based feature extraction
- L1-regularized logistic regression, tf-idf features



(a) Mean note-level AUC of regularized logistic regression over different dataset sizes. Averaged over the 5 folds, performance was optimal for each dataset size when restricted to the features with nonzero coefficients recovered by OCP.

	Fraction of training data				
Available features	**1**	**1/2**	**1/4**	**1/8**	**1/16**
OCP subset	0.864	0.860	0.847	0.808	0.786
All features	0.856	0.851	0.818	0.723	0.726
Most common	0.767	0.767	0.728	0.687	0.658
Random subset	0.740	0.747	0.727	0.639	0.634

(b) Example features that OCP selected (top) or excluded (bottom) for downstream prediction.

Selected Terms	mass, increased, decreased, stable, change, new, suspicious
Excluded Terms	discussed, imaging, left, mri, also, follow

Figure 3: Linear representation space experiment to validate assumptions of our model apply to real-world data. Quantitatively, we find downstream wins from restricting the model feature space to those found useful for the order-contrastive task. Qualitatively, the features important for the order pre-training are the same we would expect to be useful for the downstream extraction task.

Experimental Results, Part 2

- Task: predict cancer progression from patient notes via (2) BERT pre-training
- BERT base with (i) no domain-specific pre-training; (ii) FT LM: fine-tuned masked language modeling, (iii) Pt-Contrastive: a patient-level contrastive objective (identical positive sampling to OCP and PCL, but each negative is a random note of the same note type from a different patient, (iv) PCL: contrastive pre-training with PCL sampling (each negative is a random pair of notes of the same type from the same patient), (v) OCP (theirs)

Table 1: Performance of deep methods on cancer progression extraction. The first row contains the mean AUC of OCP ± its std dev. The following rows contain the mean AUC advantage of OCP over each comparison method, and the percentage of time OCP outperforms that method, across the 3 seeds and 5 folds.

	Fraction of training data				
AUC diff. (OCP Win %)	**1**	**1/2**	**1/4**	**1/8**	**1/16**
OCP AUC	0.87 ± .03	0.86 ± .04	0.84 ± .04	0.82 ± .03	0.81 ± .03
OCP – BERT	0.08 (93%)	0.12 (100%)	0.12 (100%)	0.18 (100%)	0.22 (100%)
OCP – FT LM	0.03 (80%)	0.04 (82%)	0.04 (82%)	0.08 (93%)	0.10 (89%)
OCP – Pt-Contrastive	0.03 (86%)	0.03 (77%)	0.05 (91%)	0.09 (91%)	0.12 (97%)
OCP – PCL	0.00 (53%)	0.00 (46%)	0.03 (64%)	0.03 (76%)	0.06 (87%)

Recommendations

Is it worth reading? Yes.

- Mostly straightforward paper aside from the theoretical components.
- The theoretical results are practically relevant and good to get you thinking.
- Good to consider the relationship between alternative self-supervised learning setups and the resulting representations.
- Which approach(es) makes sense for a particular problem or modality?

Is it worth implementing? Maybe.

- Effectiveness may be limited to a narrow range of downstream tasks.
- May be difficult to get data in a format that preserves time-irreversible features sufficient for extraction/learning. Consider clinical notes from a wide range of specialties, for example.