JoLT: Jointly Learned Representations of Language and Time-Series for Clinical Time-Series Interpretation (Student Abstract)

Yifu Cai, Arvind Srinivasan, Mononito Goswami, Arjun Choudhry, Artur Dubrawski

Auton Lab, School of Computer Science, Carnegie Mellon University
arvind.srini.8@gmail.com, {yifuc, mgoswami, arjuncho, awd}@andrew.cmu.edu

Abstract

Time-series and text data are prevalent in healthcare and frequently co-exist, yet they are typically modeled in isolation. Even studies that jointly model time-series and text, do so by converting time-series to images or graphs. We hypothesize that explicitly modeling time-series jointly with text can improve tasks such as summarization and question answering for time-series data, which have received little attention so far. To address this gap, we introduce JoLT to jointly learn desired representations from pre-trained time-series and text models. JoLT utilizes a Querying Transformer (Q-Former) to align the time-series and text representations. Our experiments on a large real-world electrocardiography dataset for medical time-series summarization show that JoLT outperforms state-of-the-art image captioning approaches.

Introduction

Time-series and text data are frequently recorded in routine clinical care. But unlike general text or time-series, clinical data can only be analyzed by medical professionals, who spend substantial amounts of time analyzing bio-signals, and entering summaries into electronic health records, away from direct patient care.

To cater to an ever-increasing need to effectively and efficiently interpret clinical waveforms and text data, numerous studies have been devoted to automating clinical time-series and text interpretation. However, existing studies suffer from three key limitations. First, most existing studies model time-series and text independently, even when these modalities frequently co-exist, e.g., electrocardiogram (ECG) and clinical description of findings. Second, the few studies that jointly model time-series and text are primarily rule-based, and do not offer the fluency and versatility associated with neural approaches. Third, most existing multimodal methods do not explicitly model time-series data, instead converting it to graphs or images and using graph or computer vision models, respectively.

We introduce **JoLT**, **Jo**intly Learned Representations of **L**anguage and **T**ime-series, a neural model which can generate text given time-series and textual prompts as input. We evaluate JoLT on a medical time-series summarization problem on the PTB-XL dataset, and compare it with state-of-the-art image captioning models, BLIP (Li et al. 2022) and BLIP-2 (Li et al. 2023). To the best of our knowledge, JoLT is one of the first automated ECG interpretation methods that explicitly models time-series to generate meaningful textual interpretations. Our experiments show that explicitly modeling time-series data can improve time-series summarization performance over state-of-the-art approaches pre-trained on vast amounts of data.

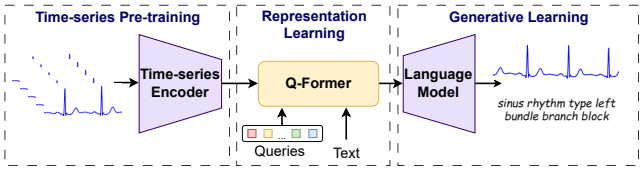


Figure 1: An overview of our training framework. We pre-train a Transformer using the masked time-series reconstruction objective to use as an encoder, and the OPT language model as the decoder. The Q-Former is trained to align time-series and text representations.

Problem Formulation and Methods

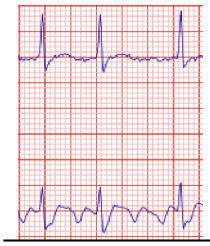
Time-series Summarization. Given a time-series $\mathcal{T} \in \mathbb{R}^{C \times L}$ of length L with C channels, our goal is to generate a textual interpretation of salient time-series features in the context of a target domain.

Model. JoLT comprises of a time-series encoder, a text decoder, and a transformer model which ties these two unimodal components together (Fig. 1). The time-series encoder is a transformer model which treats time-series subsequences as input tokens. We use the Open Pre-trained Trained (OPT) language model as a decoder. To align time-series and text representations, we leverage Querying Transformer (Q-Former) introduced by Li et al. (2023).

Pre-training Time-series Encoder. We first break the input time-series into disjoint sub-sequences called patches. A small number of patches are masked uniformly at random and then fed into the encoder, which is trained to reconstruct the masked patches using the Mean Squared Error loss.

Model	Fine-tuned	Rouge-1			Rouge-2			Rouge-L			METEOR	BLEURT
		R	P	F_1	R	P	F_1	R	P	F_1		
BLIP	×	0.006	0.010	0.007	0.000	0.000	0.000	0.006	0.010	0.007	0.025	-1.129
BLIP-2	×	0.034	0.072	0.044	0.001	0.002	0.001	0.033	0.069	0.042	0.029	-1.281
BLIP-2	✓	0.325	0.487	0.363	0.155	0.245	0.172	0.325	0.487	0.363	0.349	-0.737
JoLT	✓	**0.404**	**0.528**	**0.436**	**0.277**	**0.354**	**0.295**	**0.403**	**0.526**	**0.435**	**0.414**	**-0.502**

Table 1: JoLT outperforms zero-shot and fine-tuned state-of-the-art image captioning baselines, for the ECG interpretation task on the PTB-XL dataset. **R**, **P**, and $\mathbf{F}_1$ denote the Recall, Precision, and F_1 score. Metrics are described in the Appendix.



sinus rhythm position type normal left bundle branch block left hypertrophy possible 4.46 unconfirmed report	Ground Truth
sinus rhythm. otherwise normal ecg	Fine-tuned BLIP-2
sinus rhythm left type left bundle branch block left hypertrophy possible 4.46 unconfirmed report	JoLT

Table 2: Qualitative evaluation of the results generated by JoLT compared to fine-tuned BLIP-2. JoLT generates text summaries very similar to the ground truth, while the fine-tuned BLIP-2 got the base class correct, but incorrectly described the time-series sample.

Representation Learning. In this stage, we freeze the pre-trained encoder and train the Q-Former to learn query embeddings that capture salient time-series representations that are informative of input text. The Q-Former is trained using three objectives: (1) a *contrastive loss* to align time-series and text representations by maximizing their mutual information, (2) a *text generation loss* to train the Q-Former to generate text conditioned on input time-series, and a (3) time-series text *matching loss* for finer grained alignment between time-series and text representations.

Generative Learning. In this stage, we finally connect the frozen time-series encoder and Q-Former, with the frozen decoder, to leverage its generative capability. The query embeddings serve as *soft prompts* to guide the decoder's language generation. We train the model end-to-end using the causal language modeling loss.

Case Study: ECG Interpretation

Dataset. We conduct an experiment on the PTB-XL dataset (Wagner et al. 2020) to evaluate JoLT's ability to generate meaningful clinical interpretations from ECG waveform data. The dataset comprises of 21,837 12-lead, 10 seconds long ECG recordings collected from 18,885 patients. A subset of ECG recordings is paired with gold-standard clinical interpretation, which we use to train and fine-tune our model. The train, validation, and test sets contain 11,319, 1,636, and 1,650 samples of paired time-series and text, respectively.

Experimental Setup. We compare JoLT with state-of-the-art image captioning models BLIP and BLIP-2 as baselines. We use an ECG plotting tool[1] to transform time-series into graphical images before feeding them into these models. We evaluate both BLIP and BLIP-2 in a zero-shot setting. Since these models are not pre-trained on medical data, we also compare JoLT with BLIP-2 fine-tuned on the PTB-XL dataset. Finally, we evaluate multiple metrics that are commonly used to evaluate text generation performance.

Tables 1 and 2 summarize the results of our experiment. Below, we highlight some key observations.

Domain-specific fine-tuning is critical for clinical waveform interpretation. Poor performance of off-the-shelf BLIP-2 with respect to its fine-tuned counterpart shows the need for domain-specific fine-tuning, at least within the clinical domain. This motivates the need for publicly available large paired time-series and text datasets and models.

Explicitly modeling time-series improves summarization performance. JoLT produces textual summaries which are closer to ground truth compared to BLIP-2. We believe that the difference in performance largely stems from JoLT's ability to capture salient time-series features.

Next Steps. JoLT's encoder was pre-trained only on a small set of time-series data from the PTB-XL datasets. We believe that pre-training the encoder on a large amount of time-series data is likely to improve its performance. Furthermore, we believe that decoders trained on medical corpora would improve our clinical summaries.

Acknowledgments

This work was partially supported by the U.S. Army Research Office and the U.S. Army Futures Command under contract W911NF-20-F-0020.

References

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*.

[1] https://pypi.org/project/ecg-plot/