JoLT: Jointly Learned Representations of Language and Time-Series

Yifu Cai, Mononito Goswami, Arjun Choudhry, Arvind Srinivasan, Artur Dubrawski
Auton Lab, School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213
arvind.srini.8@gmail.com, {yifuc, mgoswami, arjuncho, awd}@andrew.cmu.edu

Abstract

Time-series and text data is prevalent in healthcare and frequently exist in tandem, for e.g., in electrocardiogram (ECG) interpretation reports. Yet, these modalities are typically modeled independently. 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 and medical question-answering approaches, and that the decoder architecture, size, and pre-training data can vary the performance on said tasks.

1 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 biosignals, 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 multi-modal 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 a state-of-the-art image captioning model, BLIP-2 [1]. 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, and can enable tasks that can be obscure for time-series and text multi-modal data, like question answering.

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

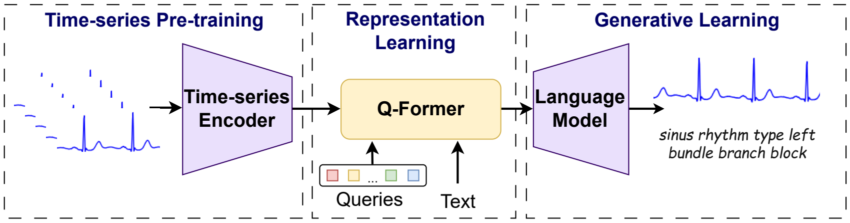

Figure 1: An overview of JoLT. Given a time-series and an optional textual prompt as input, JoLT produces text as output. We pre-train a Transformer using the masked time-series reconstruction objective to use as an encoder, and the pre-trained language model as the decoder. The Q-Former is trained to align time-series and text representations. Learnable query tokens are used to extract time-series features conditioned on textual prompts.

2 Related Work

Time-series and Text Multimodal Models. Numerous studies have explored the problem of learning multimodal representations of data, such as graphs [2], image [1], and tabular data [3], grounded in text [4]. However, the challenge of jointly modeling time-series and text data has been relatively unexplored, primarily due to the lack of large publicly available paired time-series and text (pre-) training data, i.e., there is no equivalent of LAION-5B [5] or MS-COCO [6]. On the contrary, most existing time-series datasets are domain-specific, e.g. ECG interpretation [7] or stock price variations. This is exacerbated by the fact that most existing models are either statistical or rule-based, and necessitating substantial domain expertise that does not readily transfer across different domains [8].

Clinical Text Summarization. In the healthcare domain, numerous studies have underscored the importance of developing automated text summarization systems. For instance, [9] highlights the pressing need for automated clinical report generation to alleviate the time burden on medical professionals, allowing them to focus more on patient care. Another relevant work by Harris and Zaki [10] introduced a CNN-LSTM framework designed to generate summaries of personal health data, such as heart rate, step count, and nutrient intake. Their approach demonstrated reasonably good performance in this context. However, it is worth noting that the quality of the generated summaries heavily relied on the paired training texts, which were created using rule-based methods. We expect neural methods to outperform neural methods relying on rule-based methods.

3 Problem Formulation and Methods

3.0.1 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.

3.0.2 Model.

JoLT comprises of a time-series encoder, a text decoder, and a transformer model which ties these two unimodal components together (Fig. 1) [11]. The time-series encoder is a transformer model which treats time-series sub-sequences as input tokens. Our best model uses the Open Pre-trained Trained (OPT) language model as a decoder, although we also evaluate the model with various other decoder models[12]. To align time-series and text representations, we leverage Querying Transformer (Q-Former) introduced by Li et al. [1].

Pre-training Time-series Encoder. First, we break the input time-series into disjoint sub-sequences called patches. A small percentage of these 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.

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

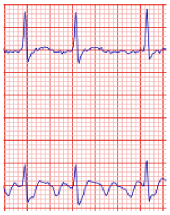	sinus rhythm position type normal left bundle branch block left hypertrophy possible 4.46 unconfirmed report	Ground Truth
	sinus rhythm. normal ecg	Fine-tuned BLIP-2
	sinus rhythm left type left bundle branch block left hypertrophy possible 4.46 unconfirmed report	JoLT (OPT-2.7B)

Table 1: Qualitative evaluation of the results generated by JoLT compared to fine-tuned BLIP-2. JoLT (with OPT-2.7B decoder) 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.

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.

4 Case Study: ECG Interpretation

Dataset. We conduct an experiment on the PTB-XL dataset [7] 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 the state-of-the-art image captioning model BLIP-2 as baseline. We use the Matplotlib package [1] to transform time-series into graphical images before feeding them into BLIP-2. We evaluate BLIP-2 in a zero-shot setting. Since the models are not pre-trained on medical data, we also compare JoLT with BLIP-2 fine-tuned on the PTB-XL dataset. We evaluate multiple metrics that are commonly used to evaluate text generation performance.

We further run ablation experiments to evaluate the impact of the decoder on our model's performance. Specifically, we evaluate different architectures (e.g. GPT-2-Large versus OPT), of different sizes (e.g. OPT-2.7B versus OPT-6.7B), and pre-trained on different data (e.g. BioGPT-Large versus GPT-2-Large). Beyond ECG interpretation, we also run preliminary experiments to explore our JoLT's ability to solve multiple choice questions conditioned on time-series, on the PTB-XL dataset. To the best of our knowledge, this unique but important problem has not been explored in prior work. We compare JoLT against BiomedCLIP[13] as a baseline.

Tables 1, 2, 3, and 4 summarize the results of our experiments. 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.

[1] https://pypi.org/project/matplotlib/

Model	Fine-tuned	Rouge-1			Rouge-2			Rouge-L			METEOR	BLEURT
		R	P	F_1	R	P	F_1	R	P	F_1		
BLIP-2	×	0.014	0.027	0.016	0.000	0.001	0.001	0.014	0.026	0.016	0.048	-1.283
BLIP-2	✓	0.217	0.343	0.227	0.100	0.193	0.107	0.215	0.341	0.225	0.202	-0.930
JoLT (OPT-2.7B)	✓	**0.404**	**0.528**	**0.436**	**0.277**	**0.355**	**0.295**	**0.403**	**0.526**	**0.435**	**0.414**	**-0.502**

Table 2: `JoLT (with OPT-2.7B decoder)` outperforms zero-shot and fine-tuned BLIP-2 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.

Decoder	Rouge-1			Rouge-2			Rouge-L			METEOR	BLEURT
	R	P	F_1	R	P	F_1	R	P	F_1		
OPT-2.7B	**0.404**	**0.528**	**0.436**	**0.277**	**0.355**	**0.295**	**0.403**	**0.526**	**0.435**	**0.414**	-0.502
OPT-6.7B	0.400	0.518	0.429	**0.277**	0.350	0.294	0.399	0.517	0.428	0.408	**-0.499**
GPT-2	0.113	0.420	0.169	0.020	0.080	0.029	0.113	0.419	0.168	0.017	-0.885
BioGPT	0.107	0.342	0.153	0.022	0.072	0.031	0.107	0.342	0.153	0.212	-1.122
BioMedLM	0.118	0.390	0.164	0.003	0.009	0.004	0.112	0.380	0.158	0.136	-1.079

Table 3: `JoLT` with OPT-2.7B decoder outperforms `JoLT` with other decoders with different pre-trained data, different sizes, and different architectures on the ECG interpretation task on the PTB-XL dataset. **R**, **P**, and $\mathbf{F_1}$ denote the Recall, Precision, and F_1 score.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F_1 Score
BiomedCLIP	0.11	**0.57**	0.11	0.02
JoLT (OPT-2.7B)	**0.54**	0.31	**0.54**	**0.4**

Table 4: Time-series conditioned multiple-choice question answering results on accuracy, precision, recall and F_1-score. `JoLT` substantially outperforms `BiomedCLIP`, when given the prompt "*Which of the following five diagnostic classes does the following ECG belong to?*". Weighted averages are measured with respect to the support for each class.

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.

Time-series and text joint modeling helps improve upon baselines for multi-class question answering. `JoLT` outperforms `BiomedCLIP` on time-series conditioned multiple choice question answering problem. However, we note that both the models perform poorly on this task. We believe that future work should carefully look to improve our model's performance on this important task, with both models overconfidently predicting the majority class.

Broken assumptions. For our ablation experiments, we hypothesized that: (1) Language models with more parameters will be better at text generation than smaller models, and (2) models pre-trained on clinical data (`BioGPT`, `BiomedLM`) will be better than those trained on general text data (`GPT-2`). However, our experiments did not support any of these hypothesis. The former can be partly explained by the fact that clinical interpretations are terse and do not require fluent large language models. We believe that further experiments are necessary to conclusively accept or reject the latter hypothesis.

5 Conclusion and Future Work

In this work, we introduced `JoLT`, a jointly model text and time-series data with a focus on ECG interpretation. We evaluated our model against state-of-the-art image captioning models in the context of clinical summarization. It's worth noting a crucial aspect of our approach: the encoder of `JoLT` was pre-trained using a relatively small set of time-series data from the PTB-XL dataset. We posit that extending this pre-training phase to include a large amount of time-series data is likely to further improve its performance. Additionally, it's important to acknowledge that the evaluation of these models presents significant challenges, as discussed in prior research [14]. Therefore, future research endeavors should aim to comprehensively and robustly assess the capabilities of such models in clinical applications.

6 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

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[2] Anthony Colas, Mehrdad Alvandipour, and Daisy Zhe Wang. GAP: A graph-aware language model framework for knowledge graph-to-text generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5755–5769, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.506`.

[3] Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. Variational template machine for data-to-text generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HkejNgBtPB`.

[4] Mandar Sharma, Ajay Gogineni, and Naren Ramakrishnan. Innovations in neural data-to-text generation: A survey, 2023.

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[7] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 2020. doi: https://doi.org/10.1038/s41597-020-0495-6.

[8] Siddharth Biswal, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Eegtotext: Learning to write medical reports from eeg recordings. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 513–531. PMLR, 09–10 Aug 2019. URL `https://proceedings.mlr.press/v106/biswal19a.html`.

[9] Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. Making effective use of healthcare data using data-to-text technology, 2018.

[10] Jonathan Harris and Mohammed J. Zaki. Towards neural numeric-to-text generation from temporal personal health data, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Transformers: Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[12] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[13] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing, 2023.

[14] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, Jul 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00879-8. URL `https://doi.org/10.1038/s41746-023-00879-8`.