Positional Encodings for Light Curve Transformers: Playing with Positions and Attention

Daniel Moreno-Cartagena [1] **Guillermo Cabrera-Vives** [1 2 3 4] **Pavlos Protopapas** [5] **Cristobal Donoso-Oliva** [3 2]
Manuel Pérez-Carrasco [2 1 4] **Martina Cádiz-Leyton** [1]

Abstract

We conducted empirical experiments to assess the transferability of a light curve transformer to datasets with different cadences and flux distributions using various positional encodings (PEs). We proposed a new approach to incorporate the temporal information directly to the output of the last attention layer. Our results indicated that using trainable PEs lead to significant improvements in the transformer performances and training times. Our proposed PE on attention can be trained faster than the traditional non-trainable PE transformer while achieving competitive results when transfered to other datasets.

1. Introduction

The Vera C. Rubin Observatory (LSST; Ivezić et al. 2019) will produce a vast number of observations every night in the sky, reaching up to 40 million events per night where the brightness or location of a source change (Sánchez-Sáez et al. 2021). The classification of this data is of utmost importance to astronomers as it allows them to acquire information about the physical characteristics and properties of astronomical objects. In recent years, the development of deep learning models has aided in categorization of light curves (Charnock & Moss 2017; Muthukrishna et al. 2019; Donoso-Oliva et al. 2021). However, light curves pose a considerable challenge as they have different distributions in each of the bands, are irregularly sampled and have varying cadences depending on the telescope at which measurements were taken (Pasquet et al. 2019; Yu et al. 2021). These peculiarities make it difficult to generate models that are sufficiently generalizable to all astronomical surveys.

Transformers, a type of deep learning model that use self-attention, have demonstrated exceptional performance on light curves (Pimentel et al. 2022; Donoso-Oliva et al. 2022; Astorga et al. 2023). In these models, temporal information is conveyed through positional encoding (PE), typically provided by sine and cosine functions with varying frequencies (Vaswani et al. 2017). However, the original proposal of PE was made in the context of text data, assuming uniform word spacing, which differs from the characteristics of astronomical time series. To tackle this challenge, we have focused on enhancing the robustness of the PE definition to effectively generalize to different astronomical surveys.

Currently, there is limited research utilizing positional encodings in transformer models to represent temporal information in light curves. Allam Jr & McEwen (2022), Donoso-Oliva et al. (2022), and Morvan et al. (2022) employed non-trainable positional encodings by directly inserting timestamps into the predefined function in Vaswani et al. (2017), achieving encouraging results compared to traditional deep learning methods. Pimentel et al. (2022) and Astorga et al. (2023) proposed a new module based on Fourier decomposition, with M harmonic components and trainable parameters, to induce temporal information. Additionally, Pan et al. (2022) utilized the rotary positional encoding proposed in Su et al. (2021) to explicitly leverage relative positions in the self-attention formulation. However, no studies have analyzed the effect of positional encoding of a transformer model within the light-curve domain.

Motivated by the above, we test different PEs, in a light curve transformer, and evaluate their performance in the reconstruction of astronomical time series and the classification of variable stars. We use the architecture proposed in Donoso-Oliva et al. (2022), which is a self-supervised light curve transformer model. We perform an empirical comparison and analysis of several PEs on astronomical surveys with different cadences. Additionally, we investigate the potential of a trainable positional encoding and propose a new approach to incorporate the temporal information.

[1]Department of Computer Science, Universidad de Concepción, Chile [2]Data Science Unit, Universidad de Concepción, Edmundo Larenas 310, Concepción, Chile [3]Millennium Nucleus on Young Exoplanets and their Moons (YEMS), Chile [4]Millennium Institute of Astrophysics (MAS), Chile [5]John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA, 02138. Correspondence to: Guillermo Cabrera-Vives <guillecabrera@inf.udec.cl>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Methods

2.1. Baseline light curve transformer model

Consider light curves of L observations, defined by a vector of fluxes $x \in \mathbb{R}^L$ and times $t \in \mathbb{R}^L$ (MJD; Modified Julian Date). A standard transformer add up the projections of the observational and temporal data into a single vector:

$$s = FFN(x) + PE(t), \tag{1}$$

where $FFN(x) \in \mathbb{R}^{L \times d_x}$ represents a Feed-Forward Network (FFN)[1] , d_x is the output size of this network, and $PE(t) \in \mathbb{R}^{L \times d_{pe}}$ is the temporal information passed through a positional encoding. To perform the addition operation, d_{pe} must equal to d_x. As proposed by Donoso-Oliva et al. (2022), the original positional encoding is expressed as a non-trainable function:

$$PE(t)_{\,2j} = \sin\left(\omega_{2j} \cdot t\right), \tag{2}$$

$$PE(t)_{\,2j+1} = \cos\left(\omega_{2j+1} \cdot t\right), \tag{3}$$

$$\omega_j = \frac{2\pi}{1000^{\frac{j}{d_{pe}}}}, \tag{4}$$

where ω_j are the angular frequencies, 1000 defines the lower bound of frequencies, t is the times vector, and $j \in [0, ..., d_{pe} - 1]$ are the dimensions of PE with a maximum of d_{pe} frequencies. Note that this PE is a slight modification of the one proposed by Vaswani et al. (2017).

The self-attention blocks receive the matrix resulting from the previous step and can be expressed as:

$$e_{ij}^{(h)} = \frac{s_i W_q^{(h)} \left(s_j W_k^{(h)}\right)^T}{\sqrt{d_k}}, \tag{5}$$

$$\alpha_{ij}^{(h)} = \frac{\exp(e_{ij}^{(h)})}{\sum_{l=1}^{L} \exp(e_{il}^{(h)})}, \tag{6}$$

$$z_i^{(h)} = \sum_{j=1}^{L} \alpha_{ij}^{(h)} \left(s_j W_v^{(h)}\right), \tag{7}$$

where $W_q^{(h)}$, $W_k^{(h)}$, and $W_v^{(h)} \in \mathbb{R}^{d_x \times d_k}$ are trainable weights matrices corresponding to the query (q), key (k), and value (v), respectively. The terms $e_{ij}^{(h)}$, $\alpha_{ij}^{(h)}$, and $z_i^{(h)}$ represent the similarity between the query and key vectors, the attention score and the output vector for each observation, respectively. Here, $h \in \{1, \ldots, H\}$ refers to the attention heads, and d_k is a hyperparameter that specifies the embedding size of the self-attention head. The output of a self-attention block considers the information of the different heads and is defined as:

$$z_i = \text{Concat}\left(z_i^{(1)}, ..., z_i^{(H)}\right) W_o, \tag{8}$$

where $W_o \in \mathbb{R}^{H \cdot d_k \times d}$ is a trainable weight matrix, and d is the dimension of the output $z \in \mathbb{R}^{L \times d}$. This output can be used as input to other multi-head attention blocks to enable the model to capture dependencies at multiple levels of abstraction. The final representation is obtained in the last block. This representation is generated by a decoder FFN that reconstructs the input fluxes $\hat{x} \in \mathbb{R}^L$ in a self-supervised objective task, by minimizing the Root Mean Square Error (RMSE) loss function. The resulting representation can serve as input for subsequent tasks, such as classification or regression.

[1]Following Donoso-Oliva et al. (2022) we use no activation function for $FNN(x)$.

2.2. Positional encodings

2.2.1. Trainable

In order to capture temporal relationships from the light curves, we replaced the angular frequencies with an embedding layer consisting of d_{pe} trainable parameters. These parameters were initialized with the predefined frequencies given in Eq. (4).

2.2.2. Fourier

We enhance the learnability and flexibility of the positional representation in Eqs. (2) and (3) by modulating it with a Multilayer Perceptron (MLP) (Li et al., 2021):

$$\hat{PE}(t) = \left(\Phi_{\text{GeLU}}\left(PE(t) \cdot W_1 + b_1\right)\right) \cdot W_2 + b_2, \tag{9}$$

where $W_1 \in \mathbb{R}^{d_{pe} \times d_m}$ and $W_2 \in \mathbb{R}^{d_m \times d_{pe}}$ are trainable parameters, $b_1 \in \mathbb{R}^{d_m}$ and $b_2 \in \mathbb{R}^{d_{pe}}$ are biases, and Φ_{GeLU} is the activation function. Here, d_m is the number of neurons in the hidden layer. In particular, W_2 projects the representation to the dimension of the input embeddings.

2.2.3. Recurrent

We followed the approach of Nyborg et al. (2022) and used a Gated Recurrent Unit (GRU; Cho et al. 2014) to incorporate temporal dependencies at time steps t expressed with the baseline positional encoding shown in Eqs. (2) and (3):

$$o(t) = \text{GRU}\left(PE(t)\right), \tag{10}$$

$$\hat{PE}(t) = o(t) \cdot W_p + b_p, \tag{11}$$

where $o(t) \in \mathbb{R}^{L \times d_{pe}}$ is the output of the GRU at each time step, $W_p \in \mathbb{R}^{d_{pe} \times d_{pe}}$ is a trainable weight matrix and $b_p \in \mathbb{R}^{d_{pe}}$ is the trainable bias.

2.2.4. Tupe-A

We follow the approach employed in Ke et al. (2020) for natural language processing (NLP) and separate the mixed

correlations produced between observational and temporal information in the attention matrix by redefining Eqs. (5) and (7). To represent the queries and keys of the temporal information expressed by Eqs. (2) and (3), we introduce new parameters $U_q, U_k \in \mathbb{R}^{d_{pe} \times d_k}$, respectively:

$$e_{ij}^{(h)} = \frac{FFN(x)_i W_q^{(h)} \left(FFN(x)_j W_k^{(h)}\right)^T}{\sqrt{d_k}} + \frac{PE(t)_i U_q^{(h)} \left(PE(t)_j U_k^{(h)}\right)^T}{\sqrt{d_k}}, \tag{12}$$

$$z_i^{(h)} = \sum_{j=1}^{L} \alpha_{ij}^{(h)} (x_j W_v^{(h)}). \tag{13}$$

For efficiency, we share these new parameters across different multi-head attention blocks (Ke et al., 2020).

2.2.5. CONCAT

Following the same idea and aiming to minimize the noise generated in the attention matrix due to mixed correlations between observational and temporal information, we concatenated them in separate orthogonal spaces:

$$s = [FFN(x) \,||\, PE(t)]\,. \tag{14}$$

In this case, we utilize the trainable PE described in subsection 2.2.1.

2.2.6. PE ON ATTENTION (PEA)

To avoid mixing information, we propose incorporating positional encoding directly into the final representation obtained from the last multi-head attention block:

$$\hat{z} = z + PE(t), \tag{15}$$

where $PE(t)$ is expressed by the baseline positional encoding shown in Eqs. (2) and (3) as a non-trainable function. Here, the multi-head attention block takes only the observational information $FFN(x)$ to compute the attention.

3. Experiments

3.1. Data description

For the pretraining stage, we utilized the unlabeled MACHO light curves dataset (Alcock et al., 2000) and excluded curves exhibiting noisy behavior[2]. The dataset comprised a total of $1,529,386$ light curves in the R-band, with a median cadence of 1.00 days (refer to Appendix A for the cadence distribution). Subsequently, we evaluated the performance of the pretrained transformers on the classification task using a subset of 500 objects per class from the MACHO labeled survey (*Full* hereafter, Cutri et al., 2003). To isolate the effect of cadence at this task, we simulated three datasets from the MACHO labeled subset by modifying the cadence of the light curves. Specifically, we removed observations from the light curves at rates of $3/4$, $1/2$, and $1/4$, respectively. At a rate of $3/4$, we removed the last observation out of four, while at a rate of $1/4$, the last three observations were removed. To account for external factors, such as changes in the band distribution, we also tested the pretrained transformers on OGLE-III (Udalski, 2004), which contains $358,288$ I-band light curves, and ATLAS (Heinze et al., 2018), which contains $422,630$ orange-band light curves. Specifically, we used a subsample of 500 objects per class from each labeled data subsets to consider the scenario where we have few labeled data. The flux distribution, cadence distribution and classes of each labeled subset can be found in Appendix A.

[2]We defined noise as points in the light curve with $|Kurtosis| > 10$, $|Skewness| > 1$, and a flux error > 0.1.

3.2. Training details

We ran the experiments on a Nvidia RTX A5000 GPU, employing two multi-head attention blocks with $H = 4$ heads and $d_k = 64$ neurons. The model dimensions were set at $d = d_x = d_{pe} = 256$ and $d_l = 64$, with the exception of Concat PE, which employed $d_x = d_{pe} = 128$. Light curve windows with a maximum length of $L = 200$ were considered. For light curves that exceeded this length, subsequent time windows were sampled, beginning from a random position. For light curves with fewer than L observations, zero values were padded at the end. Each generated window was subtracted from its observational and temporal mean, creating flux and time vectors with zero mean.

For the pretraining stage, we followed the strategy used in Devlin et al. (2018), masking a percentage of the observations in the light curves. Specifically, we selected 50% of the observations in each light curve for evaluating the reconstruction of the flux $\boldsymbol{x}$ in the loss function. Within this percentage, we masked 30% of the observations, replaced 10% with random values, and left the remaining 10% of observations visible. We used early stopping with patience of 40 epochs on the validation loss. The Adam optimizer (Kingma & Ba, 2014) was used with a learning rate of 10^{-5} and a batch size of $2,000$.

For the classification task, we used two hidden layers of 256 Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber 1997) units followed by a MLP with a softmax activation function. The dimension of this output layer depends on the number of classes to be classified. We also divided the training and validation sets into 3 folds with an $80/20$ ratio, respectively. We used early stopping with a patience of 20 epochs on the validation loss and the Adam optimizer with

Table 1. Performance of different positional encodings in the pretraining stage and classification task.

PE TYPE	MACHO UNLAB.		MACHO LAB.				OGLE	ATLAS
			FULL	3/4	1/2	1/4		
	RMSE	TIME (EPOCHS)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)
BASELINE	.170	6D 14H (523)	71.6 ± 1.9	69.2 ± 1.9	66.2 ± 1.9	63.3 ± 1.5	71.3 ± 1.1	65.8 ± 1.4
TRAINABLE	**.169**	2D 13H (202)	72.9 ± 2.1	72.3 ± 1.0	$\mathbf{71.0 \pm 1.0}$	$\mathbf{69.0 \pm 0.5}$	74.9 ± 1.4	65.4 ± 1.8
FOURIER	.170	1D 20H (142)	73.0 ± 1.1	70.2 ± 1.9	67.8 ± 0.9	62.9 ± 2.0	72.0 ± 0.8	$\mathbf{69.6 \pm 0.1}$
RECURRENT	.197	0D 16H (048)	67.1 ± 1.8	63.5 ± 2.5	59.7 ± 1.9	54.6 ± 1.3	70.7 ± 1.1	68.3 ± 0.9
TUPE-A	.219	0D 17H (084)	67.3 ± 1.6	66.1 ± 1.4	64.9 ± 1.0	60.8 ± 0.9	71.0 ± 1.0	67.5 ± 0.9
CONCAT	.170	3D 01H (237)	$\mathbf{73.4 \pm 1.1}$	$\mathbf{73.1 \pm 1.7}$	70.9 ± 1.7	$\mathbf{69.0 \pm 1.8}$	74.5 ± 1.3	68.1 ± 0.6
PEA	.199	0D 17H (058)	69.7 ± 0.9	68.9 ± 1.8	68.0 ± 1.0	65.5 ± 2.5	$\mathbf{76.3 \pm 1.2}$	66.9 ± 1.0

a learning rate of 10^{-4} and a batch size of 512.

3.3. Results

Table 1 provides the evaluation of a transformer pretrained from scratch using different positional encodings on both simulated and real datasets. We pretrained each transformer on the unlabeled MACHO data and evaluated the reconstruction of the observational information in terms of RMSE and training time. We then used the generated representation for training the classification layers on each labeled dataset and evaluated its performance in terms of F1-score. Our baseline is the fixed positional encoding described in subsection 2.1.

During pretraining, the Trainable PE demonstrated a slight improvement in terms of the reconstruction RMSE and a significant reduction in training time when compared to the baseline. The Fourier and Concat PE matched the performance of the baseline while needing less computational resources. The Fourier PE showed the best performance in terms of both training time and reconstruction performance. Recurrent, Tupe-A, and PEA did not outperform the Baseline in RMSE terms, but converged in less than a day of training. Learning curves are shown in Appendix B.

Since we are analyzing which PE can generate a better representation during the pretraining stage, we keep the transformer, including the PE, fixed when training for the classification task. We start by evaluating the performance of the transformers on the MACHO labeled datasets considering the effect of the change in cadence. The Trainable PE outperformed the Baseline for all cadences. In particular, the degradation of results for sparser light curves is less severe with the Trainable PE than with the non-trainable one. Similarly, Fourier PE performs better than the baseline on three out of four datasets. However, the degradation of results, as evaluated on the $1/2$ and $1/4$ cadences, was similar to the baseline and worse than the Trainable PE. The Recurrent and Tupe-A PE show worse classification performance than the baseline. The Concat PE outperformed the Baseline and achieved three of the four best performances in terms of F1-score. Its degradation was minimal for sparser light curves, and close to the Trainable PE. Finally, PEA outperformed the baseline for cadences of $1/2$ and $1/4$, but did worse for the full and $3/4$ cadences.

Upon adding changes in the fluxes distributions using OGLE and ATLAS, we observe that the baseline exhibits inferior overall F1-score performance. The trainable PE model outperforms the baseline in OGLE and demonstrated a similar performance as the baseline for ATLAS. The Fourier PE model showed superior performance in both astronomical surveys with respect to the baseline. However, it did not outperform the Trainable PE model in OGLE. In particular, the Fourier PE model obtained the best F1-score in ATLAS. Similarly, the Recurrent and Tupe-A PE models outperformed the baseline model in ATLAS and while exhibiting a similar performance in OGLE. Finally, the Concat PE and PEA models outperformed the baseline in both OGLE and ATLAS, with the latter obtaining the highest F1-score among all the PE on OGLE.

The separation of temporal and observational information into orthogonal spaces (Concat PE) results in better classification performance on all datasets, yielding the best average F1-score overall. Recall that both the Trainable and Concat PE use the same trainable PE: the first add the PE to a vectorized representation of the fluxes, while the second concatenates these vector. Both of these PEs show a small degradation in results for the MACHO datasets with different cadences, implying that they allow for a better representation of temporal information.

In terms of training time, all the trainable PEs and the proposed PEA exhibit reduced pretraining time compared to the baseline. Out of the three models that take less than one day to train (Recurrent, Tupe-A, and PEA), the best classification results for the different MACHO cadence datasets are achieved by our proposed PEA. At the same time, PEA outperforms the Recurrent and Tupe-A PEs when transfered to

OGLE, while the three of them achieve similar classification results on the ATLAS dataset (less than 1.6 sigma). This is of particular importance when training large light curve models with massive datasets for next generation surveys such as the LSST.

4. Conclusion

In this work, we have evaluated the transferring potential of a light curve transformer to datasets with different cadences and flux distributions. Our results have demonstrated that using a trainable positional encoding offers advantages over a non-trainable PE baseline, in terms of both model performance and computational efficiency. Additionally, we have highlighted the benefits of separating observational and temporal information within the attention matrix and proposed a new approach for incorporating temporal information directly into the output of the last attention layer. At the same time, our proposed method trains faster than the baseline while achieving competitive classification performances.

Acknowledgements

The authors acknowledge support from the National Agency for Research and Development (ANID) grants: FONDECYT regular 1231877 (DMC, GCV, MCL); Millennium Science Initiative Program – NCN2021_080 (GCV, CDO) and ICN12 009 (GCV, MPC).

References

Alcock, C., Allsman, R., Alves, D. R., Axelrod, T., Becker, A. C., Bennett, D., Cook, K. H., Dalal, N., Drake, A. J., Freeman, K., et al. The macho project: microlensing results from 5.7 years of large magellanic cloud observations. *The Astrophysical Journal*, 542(1):281, 2000.

Allam Jr, T. and McEwen, J. D. Paying attention to astronomical transients: Introducing the time-series transformer for photometric classification. 2022.

Astorga, N., Reyes, I., Cabrera, G., Förster, F., Huijse, P., Arredondo, J., Moreno-Cartagena, D., Muñoz-Arancibia, A., Bayo, A., Catelan, M., et al. Atat: Astronomical transformer for time series and tabular data. 2023.

Charnock, T. and Moss, A. Deep recurrent neural networks for supernovae classification. *The Astrophysical Journal Letters*, 837(2):L28, 2017.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Cutri, R. e., Skrutskie, M., Van Dyk, S., Beichman, C., Carpenter, J., Chester, T., Cambresy, L., Evans, T., et al. Vizier online data catalog. *II/246*, 3, 2003.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Donoso-Oliva, C., Cabrera-Vives, G., Protopapas, P., Carrasco-Davis, R., and Estévez, P. A. The effect of phased recurrent units in the classification of multiple catalogues of astronomical light curves. *Monthly Notices of the Royal Astronomical Society*, 505(4):6069–6084, 2021.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vardhan, H., et al. Astromer: A transformer-based embedding for the representation of light curves. *arXiv preprint arXiv:2205.01677*, 2022.

Heinze, A., Tonry, J. L., Denneau, L., Flewelling, H., Stalder, B., Rest, A., Smith, K. W., Smartt, S. J., and Weiland, H. A first catalog of variable stars measured by the asteroid terrestrial-impact last alert system (atlas). *The Astronomical Journal*, 156(5):241, 2018.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.

Ke, G., He, D., and Liu, T.-Y. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, Y., Si, S., Li, G., Hsieh, C.-J., and Bengio, S. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021.

Morvan, M., Nikolaou, N., Yip, K. H., and Waldmann, I. Don't pay attention to the noise: Learning self-supervised representations of light curves with a denoising time series transformer. *arXiv preprint arXiv:2207.02777*, 2022.

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., and Hložek, R. Rapid: early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002, 2019.

Nyborg, J., Pelletier, C., and Assent, I. Generalized classification of satellite image time series with thermal positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1392–1402, 2022.

Pan, J., Ting, Y.-S., and Yu, J. Astroconformer: Inferring surface gravity of stars from stellar light curves with transformer. *arXiv preprint arXiv:2207.02787*, 2022.

Pasquet, J., Pasquet, J., Chaumont, M., and Fouchez, D. Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, 627:A21, 2019.

Pimentel, Ó., Estévez, P. A., and Förster, F. Deep attention-based supernovae classification of multiband light curves. *The Astronomical Journal*, 165(1):18, 2022.

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., Förster, F., Eyheramendy, S., Elorrieta, F., Bauer, F., Cabrera-Vives, G., Estévez, P., Catelan, M., et al. Alert classification for the alerce broker system: The light curve classifier. *The Astronomical Journal*, 161(3):141, 2021.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Udalski, A. The optical gravitational lensing experiment. real time data analysis systems in the ogle-iii survey. *arXiv preprint astro-ph/0401123*, 2004.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yu, C., Li, K., Zhang, Y., Xiao, J., Cui, C., Tao, Y., Tang, S., Sun, C., and Bi, C. A survey on machine learning based light curve analysis for variable astronomical sources. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1425, 2021.

A. Dataset description.

In this section, we provide essential information to analyze the flux, cadence, and class distributions of various astronomical surveys employed in this study. Figure 1 illustrates the flux distribution across the different datasets. It shows the similarity in flux distributions among the various sets of labeled MACHO (full, $3/4$, $1/2$, and $1/4$) and the disparities in flux distributions between the OGLE, ATLAS, and labeled MACHO datasets. Figures 2 and 3 display the cadence distributions for the unlabeled dataset and the labeled datasets, respectively. Key statistical measures are provided to understand the sampling frequency of observations within the light curves. The unlabeled MACHO dataset, which served as the pretraining data for the transformers, exhibits a median of 1.00 and a mean of 2.96 with a standard deviation of 17.19. These values indicate the time gap between successive observations and the temporal separation between groups of observations in the light curves. In particular, the unlabeled MACHO dataset demonstrates a smaller time gap compared to the labeled MACHO dataset, while also exhibiting a higher standard deviation, implying greater temporal separation between groups of observations. The labeled MACHO-derived datasets (3/4, 1/2, and 1/4) exhibit an increase in both the median and standard deviation as light

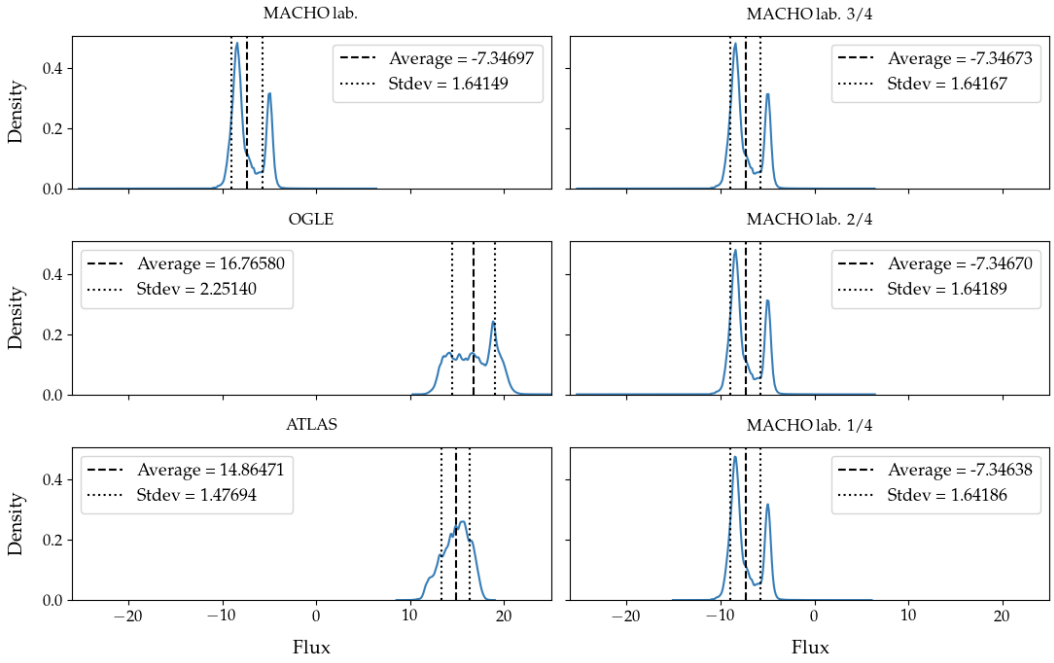

Figure 1. Flux distribution of the different data subsets.

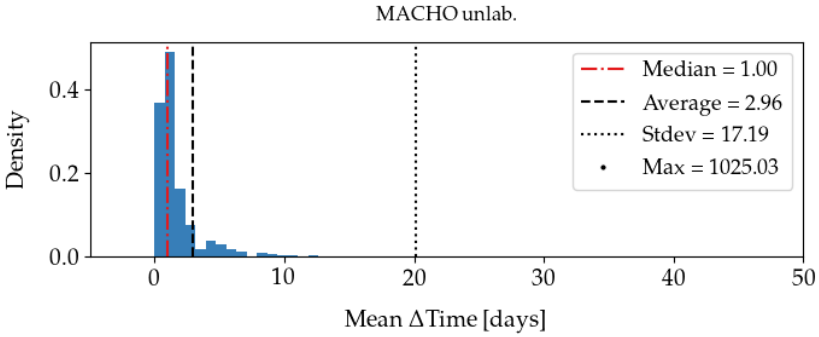

Figure 2. Unlabeled MACHO cadence distribution.

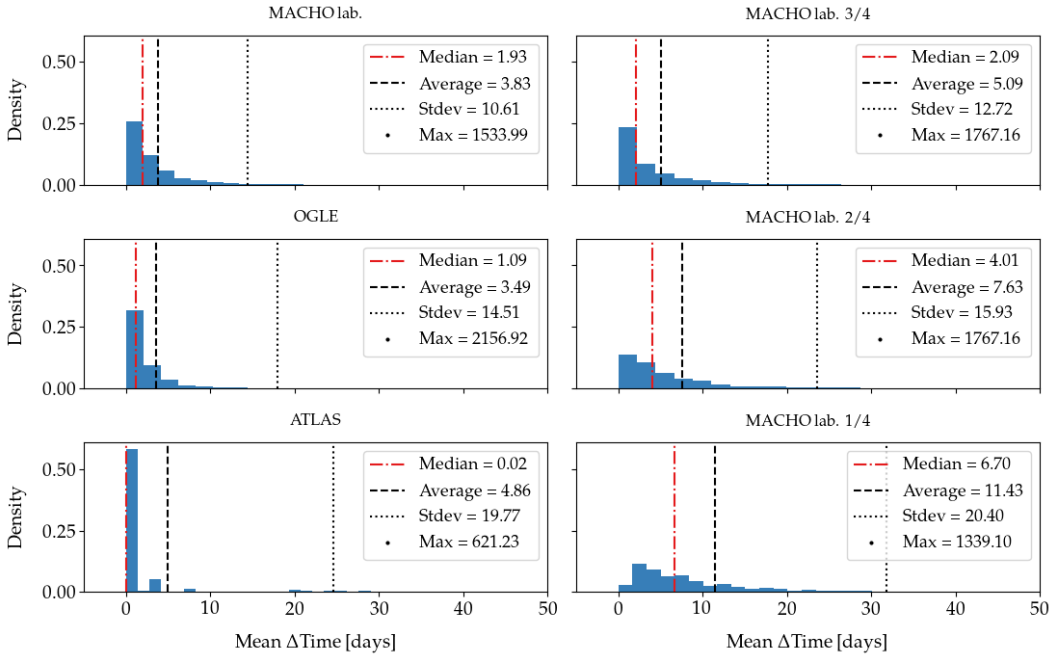

Figure 3. Cadence distribution of the different data subsets.

curve observations are reduced. Additionally, they show distinctions compared to both the labeled and unlabeled MACHO datasets. Regarding OGLE, its cadence displays similarities with that of the unlabeled MACHO dataset. However, ATLAS exhibits more pronounced time gaps between groups of observations. The mean is influenced by the standard deviation, while the median indicates that the observations are taken at short time intervals.

Table 2 displays the classes used for each of the datasets. MACHO labeled has six classes, OGLE has ten classes, and ATLAS has four classes. The modified cadence datasets maintain the same number of classes as the MACHO labeled dataset.

Table 2. Labels from each of the datasets used in the classification task.

TAG	MACHO LAB.	OGLE	ATLAS
EC	ECLIPSING BINARY	ECLIPSING BINARY	-
ED	-	DETACHED BINARY	DETACHED BINARY
ESD	-	SEMI-DETACHED BINARY	-
MIRA	-	MIRA	MIRA
OSARG	-	SMALL-AMPLITUDE RED GIANT	-
RRAB	RR LYRAE TYPE AB	RR LYRA TYPE AB	
RRC	RR LYRAE TYPE C	RR LYRAE TYPE C	
DSCT	-	DELTA SCUTI	PULSE
CEP_0	CEPHEID TYPE I	CEPHEID	
CEP_1	CEPHEID TYPE II		
SRV	-	SEMI-REGULAR VARIABLE	-
LPV	LONG PERIOD VARIABLE	-	-
CB	-	-	CLOSE BINARIES

B. Pretraining learning curves.

In this section, we present the learning curves on the validation set for the proposed PEA and the positional encodings that achieved superior RMSE during pretraining. Figure 4 illustrates the pretraining of transformers using the same hyperparameters. The y-axis represents the mean value of RMSE with a 4-step window, and the x-axis represents the number of epochs displayed on a logarithmic scale. It is evident that trainable positional encodings such as Trainable, Fourier, and Concat PE achieved comparable RMSE to the baseline with significantly fewer epochs in pretraining (38.6%, 27.2%, and 45.3% of baseline epochs, respectively). In contrast, the PEA method initially obtained a higher RMSE than the baseline, but it achieved an average RMSE of 0.204 earlier than the baseline by utilizing only 57.7% of the epochs.

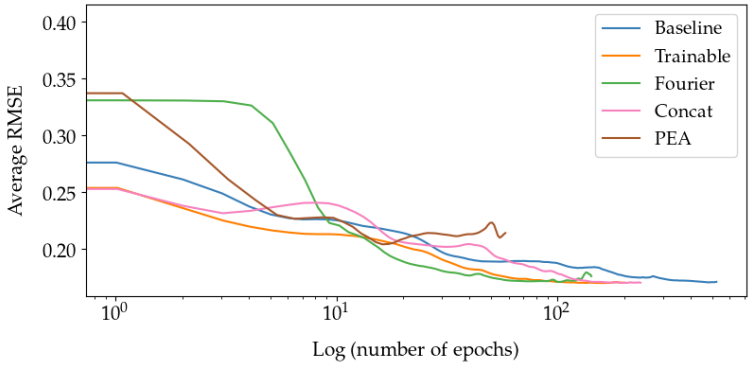

Figure 4. Validation loss in pretraining stage.