ASTROMER Model: Formal Pseudocode

Shivam Tyagi

Positional Encoding

Algorithm 1 Positional Encoding

Input: $t \in \mathcal{R}$, observation times in MJD
Output: $z \in \mathcal{R}^{d_{pe}}$, positional encoding vector representation
Parameters: d_{pe} dimension of positional embedding, l_{max} maximum length for observation times, base = 1000

1. $\forall i : z[2i-1] \leftarrow \sin(\frac{t}{\text{base}^{2i/d_{pe}}})$
2. $\forall i : z[2i] \leftarrow \cos(\frac{t}{\text{base}^{2i/d_{pe}}})$
3. **return** z

Feed-Forward Network (FNN) for Magnitudes

Algorithm 2 Feed-Forward Network for Magnitudes

Input: $e_t \in \mathcal{R}^{d_e}$, vector representation of the magnitude
Output: $z_t \in \mathcal{R}^{d_{pe}}$, transformed magnitude vector representation
Parameters: $W_{mlp} \in \mathcal{R}^{d_{pe} \times d_e}$, $b_{mlp} \in \mathcal{R}^{d_{pe}}$, MLP parameters

1. $z_t \leftarrow W_{mlp} \cdot e_t + b_{mlp}$
2. **return** z_t

Multi-Head Attention (MHA)

Encoder

Decoder (Pre-Training Phase)

Algorithm 3 Multi-Head Attention

Input: $E \in \mathcal{R}^{d_e \times l_{max}}$, embedded vectors
Output: $A \in \mathcal{R}^{d_e \times l_{max}}$, attention output
Parameters: Attention block parameters
$W_q \in \mathcal{R}^{d_{attn} \times d_e}$, $W_k \in \mathcal{R}^{d_{attn} \times d_e}$, $W_v \in \mathcal{R}^{d_{out} \times d_e}$, $W_o \in \mathcal{R}^{d_{out} \times d_e}$

1. $\forall t : q_t \leftarrow W_q \cdot E_t$
2. $\forall t : k_t \leftarrow W_k \cdot E_t$
3. $\forall t : v_t \leftarrow W_v \cdot E_t$
4. $\forall t, t' : \alpha_{t,t'} \leftarrow \frac{\exp(q_t^\top k_{t'} / \sqrt{d_{attn}})}{\sum_u \exp(q_t^\top k_u / \sqrt{d_{attn}})}$
5. $\forall t : a_t \leftarrow \sum_{t'} \alpha_{t,t'} \cdot v_{t'}$
6. $A \leftarrow W_o \cdot A$
7. **return** A

Algorithm 4 Encoder

Input: $z \in \mathcal{R}^{d_{pe} \times l_{max}}$, positional encoding
Input: $e_t \in \mathcal{R}^{d_e}$, vector representation of the magnitude
Output: $E \in \mathcal{R}^{d_e \times l_{max}}$, encoded representations
Parameters: L, number of encoder layers, H, number of attention heads

1. $E \leftarrow z + \mathrm{FNN}(e_t)$
2. $\forall l \in [1, L] : E \leftarrow \mathrm{LayerNorm}(E + \mathrm{MultiHeadAttention}(E))$
3. $\forall l \in [1, L] : E \leftarrow \mathrm{LayerNorm}(E + \mathrm{FeedForward}(E))$
4. **return** E

Algorithm 5 Decoder (Pre-Training Phase)

Input: $E \in \mathcal{R}^{d_e \times l_{max}}$, encoded representations
Output: Reconstructed Magnitudes $\in \mathcal{R}^{1 \times l_{max}}$
Parameters: $W_d \in \mathcal{R}^{d_e \times d_{out}}$, $b_d \in \mathcal{R}^{d_{out}}$

1. $\forall t : \mathrm{Reconstructed}_t \leftarrow W_d \cdot E_t + b_d$
2. **return** Reconstructed Magnitudes

Self-Supervised Pre-Training Task

Algorithm 6 Self-Supervised Pre-Training Task

Input: Magnitudes $\in \mathcal{R}^{1\times l_{max}}$, original magnitudes
Output: Loss, reconstruction loss for masked magnitudes
Parameters: p_{mask} (probability of masking)

1. $\forall t$: $\text{MaskedMagnitudes}_t \leftarrow \text{Mask}(Magnitudes_t, p_{mask})$
2. $E \leftarrow \text{Encoder}(\text{PositionalEncoding}(Times), \text{FeedForward}(\text{MaskedMagnitudes}))$
3. $\text{Reconstructed} \leftarrow \text{Decoder}(E)$
4. $\text{Loss} \leftarrow \text{MSE}(\text{Reconstructed}, Magnitudes)$
5. **return** Loss

Method	Complexity	Memory Usage	Accuracy	Best Use Case	Drawbacks
Standard Self-Attention	$O(n^2)$	High	High (Exact Attention)	Short to moderate sequence lengths	Unscalable for long sequences
FlashAttention	$O(n^2)$	Reduced (Optimized for GPUs)	High (Exact Attention)	Long sequences in modern hardware	Requires specialized hardware (e.g., modern GPUs)
Linformer	$O(n)$	Low	Medium (Approximate Attention)	Long sequences with memory constraints	Approximation error in attention weights
Performer	$O(n)$	Low	High (Near Exact Attention)	Long sequences, large datasets	Kernel-based approximation complexity
Reformer	$O(n \log n)$	Low	Medium (Approximate Attention)	Very long sequences, sparse attention	Some loss of accuracy, only effective for sparse data
Grouped Query Attention (GQA)	$O(n^2)$	Reduced (due to query grouping)	Medium (Attention on Groups)	Tasks with repeated queries	Loss of fine-grained query details