GLA: Global-Local self-Attention for Enhancing Fine-Tuning of Transformers on Temporal Structured Health Data

Anonymous authors
Paper under double-blind review

Abstract

Pre-training transformer models on self-supervised tasks and fine-tuning them on downstream tasks, even with limited labeled samples, have achieved state-of-the-art performance across various domains. However, learning effective representations from complex temporal structured health data and fine-tuning for health-related risk predictions remains challenging. While self-attention mechanisms excel in capturing relationships within sequences, they may struggle to adequately model both long-range dependencies (global attentions) and short-range dependencies (local attentions) within a sequence of events. Addressing this limitation typically involves expensive enhancements to the pre-training process. In this work, we propose a novel method called Global-Local self-Attention (GLA) to augment pre-trained models during the fine-tuning phase. Our approach encourages the self-attention mechanism to effectively capture both long and short dependencies within input sequences simultaneously. This is achieved by introducing noise to the self-attention and then convolving it with a 2D Gaussian kernel. The first term encourages attention between distant events in the input sequence (global attention), while the second term promotes attention to local events. With GLA, we observe enhanced model performance on downstream tasks. Furthermore, our method sheds light on the model's ability to learn complex global and local relations within a sequence of medical events, providing valuable insights into its behavior within the attention mechanism.

1 Introduction

Foundation models, deep neural networks pre-trained on broad unlabeled data using self-supervised methods, have significantly impacted various aspects of our lives, including law, healthcare, education, and more Bommasani et al. (2021); Guo et al. (2023); Wornow et al. (2023). These models typically acquire general knowledge about the data through pre-training a variant of the transformer network on a self-supervised task like Masked Language Model (MLM), and then adapt this knowledge to downstream tasks with only a few labeled samples during the fine-tuning process.

Pre-training transformers have been employed with various self-supervised objectives and domains. Common objectives include corrupted text reconstruction tasks like MLM Devlin et al. (2018); Lewis et al. (2019); Lan et al. (2019) and standard language models such as next-word prediction Radford et al. (2019); Brown et al. (2020), which have been extensively utilized Liu et al. (2023). These models typically adopt a backbone architecture inspired by the multi-head attention mechanism in transformers Vaswani et al. (2017), known for its effectiveness in modeling complex interaction between events (tokens) in a sequence (text). These foundation models have been pre-trained on different domain data, such as general text Lan et al. (2019); Radford et al. (2019) and structured temporal health data as sequences of events Li et al. (2020); Rasmy et al. (2021); Pang et al. (2021); Amirahmadi et al. (2024a).

Modeling Electronic Health Records (EHRs) trajectories presents a critical opportunity for predicting health-related outcomes, offering benefits like early intervention, cost reduction, and improved public health. This field has attracted significant attention from deep learning researchers Xiao et al. (2018); Amirahmadi et al. (2023); Boll et al. (2024). Typically, healthcare specific foundation models are pre-trained on extensive, publicly available, unlabeled EHR data, and adapting these models through fine-tuning consistently demonstrates superior performance across various tasks Li et al. (2022); Amirahmadi et al. (2024b); Ren et al. (2021).

However, Adapting these pre-trained models for modeling temporal structured healthcare data comes with its challenges. Researchers have explored the effectiveness of self-attention architectures in capturing complex long-term and short-term dependencies within healthcare events (Li et al., 2020; Rasmy et al., 2021). While self-attention architectures offer direct connections within event sequences and global and local attending abilities, they still face limitations in capturing the full capacity of self-attentions for modeling intricate dependencies (Choi et al., 2020; Zhu & Razavian, 2021; Amirahmadi et al., 2024a). Challenges such as data scarcity due to privacy concerns, extensive sparsity in datasets, and imbalanced labeled samples for specific diseases further compound the issue. To address the lack of attention challenge, researchers have experimented with various approaches, including initializing self-attentions with domain knowledge (Choi et al., 2020), employing variational autoencoders (Zhu & Razavian, 2021), and incorporating auxiliary tasks during pre-training (Pang et al., 2021; Amirahmadi et al., 2024a; Ren et al., 2021). Figure 1 illustrates how these different solutions impact attention behaviors. However, these methods often come with computational costs and needs extra effort for implementation and design. In this study, we propose a simple two-step augmentation method aimed at encouraging self-attention to learn and acquire complex dependencies within event sequences during the fine-tuning step. This method does not alter the computational graph, making it applicable to any pre-trained network with minimal effort.

The primary focus of our work is to enhance the performance of pretrained transformers during the fine-tuning stage in healthcare applications, specifically through on-the-fly augmentation with GLA (Global-Local self-atttentin Augmentation) without altering its computational graph. This approach is comparable to recent augmentation methods like Neftune, which also operate within the pretrained transformer framework. Importantly, our method maintains flexibility and is applicable across any domain utilizing transformers, a quality that ensures broad applicability beyond the specific healthcare datasets we used.

The main contributions are summarized as follows:

1. We proposed a simple augmentation method for self-attention to encourage global and local attentions without altering the computational graph during the fine-tuning step.
2. We conducted several evaluations on various downstream tasks, examining the effect of the novel method on model performance, model robustness with limited training samples,and the balance of attention distribution between distant and nearby events. Our results demonstrate how it improves the performance of pre-trained transformers.

2 Preliminary

2.1 Transformer encoder and self-attention

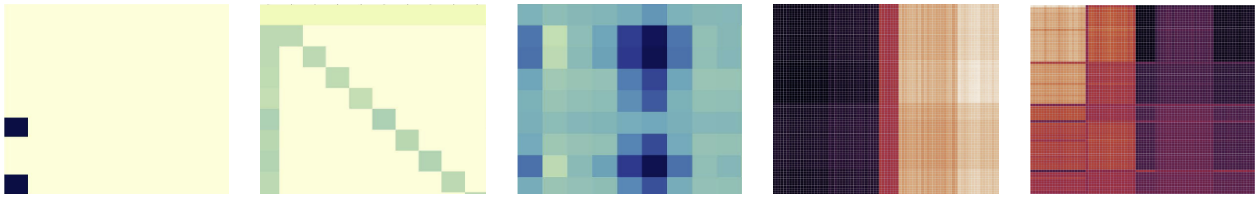
The core back-bone of transformers encoder is the multi-head self-attention. Each self-attention head is:

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V, \tag{1}$$

$$A_h = \mathrm{softmax}(\frac{Q_h K_h^T}{\sqrt{d_k}}) \tag{2}$$

$$H_h = \text{Self-attention}(X) = A_h V_h \tag{3}$$

Where, $Q, K \in \mathbb{R}^{n \times d_k}$ and $V \in \mathbb{R}^{n \times d_v}$ and n is the length of input sequence and d_k and d_v are dimenssion of Key and Value. A_h is the attention score matrix and each $A_{i,j}$ indicates how much attention token x_i put



(a) Transformer (b) GCT (c) VGNN (d) Pre-trained Transformer (e) TOO-Bert

Figure 1: Visualization of attention score patterns for five different models from previous studies. (a) A randomly initialized transformer displaying a poor structure Zhu & Razavian (2021). (b) GCT model exhibiting improved structures through the incorporation of domain knowledge Choi et al. (2020). (c) VGNN revealing complex structures with a variational regularization encoder Zhu & Razavian (2021). (d) A transformer pre-trained on MLM predominantly attending to the time distance between events and the final outcome Amirahmadi et al. (2024a). (e) Pre-training a transformer on MLM and trajectory order prediction showcasing a complex structural pattern Amirahmadi et al. (2024a). Panels (a), (b), and (c) are adapted from Zhu & Razavian (2021), trained for AD prediction, while panels (d) and (e) are adapted from Amirahmadi et al. (2024a), trained for HF prediction.

on x_j. Transformer encoders, is built on concatenation of $\mid h \mid$ number attention heads in parallel, so each one has its own weights. Then, the concatenation is projected:

$$\text{MultiHead}(X) = \text{Concat}(H_1...., H_{|h|})W^O \tag{4}$$

Where, $W^O \in \mathbb{R}^{|h| \times d_v}$ Multiple self-attention heads in parallel, help the model to attend to information from different representation subspaces (Vaswani et al., 2017; Hao et al., 2021).

2.2 Pre-training, fine-tuning

Pretraining typically involves the model acquiring general knowledge, which is then used to initialize the final network. Subsequently, the final network adjusts these weights to obtain optimized weights for specific downstream tasks Chen et al. (2021). This approach has been extensively utilized for adapting foundation models to downstream tasks Lan et al. (2019); Liu et al. (2023).

3 Related works

Regularization and augmentation methods have been vastly used to enhance model fine-tinning. Regularization methods are primarily developed to prevent overfitting and leverage acquired prior knowledge effectively. Various approaches have been proposed for fine-tuning regularization, including adding L^2 norm regularization between pretrained and downstream model parameters (Xuhong et al., 2018), knowledge distillation (Yim et al., 2017), and utilizing pretrained labels to regulate the fine-tuning process (You et al., 2020). Attention maps have been incorporated into convolutional neural networks for regularization purposes (Li et al., 2019; Zagoruyko & Komodakis, 2016), while (Zhou et al., 2023) enforced classification error of the downstream tasks' head on the pre-trained feature distributions. Additionally, (Kim et al., 2023) prioritized discriminative information during the fine-tuning step to regularize the fine-tuned model. (Zehui et al., 2019) proposed DropAttention, a self-attention specialized regularization method that masks the attention score matrix randomly and expands the mask with the span length. (Wu et al., 2023) advanced self-attention generalization by introducing an adversarial structural bias to the attention score matrix, demonstrating that naively masking the attention score can improve transformer performance, albeit with the complexity and training overhead of adversarial training being a notable issue.

Moreover, researchers have also focused on enhancing the performance and generalizability of various models by augmenting noise into the model during fine-tuning. Zhu et al. (2019) improved the generalization of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) by adding adversarial perturbation noise to the word embeddings. This approach has been extended to graph neural networks by Kong et al. (2022), enhancing generalizability to out-of-distribution samples. Wornow et al. (2023) added normal noise to the latent space representation of an encoder-decoder architecture to improve image captioning. Additionally, Jain et al. (2023) applied uniform noise to the embedding vectors during fine-tuning, defined within a range of $[-\alpha/\sqrt{Ld}, \alpha/\sqrt{Ld}]$, where L is the sequence length, d is the embedding dimension, and α is a tunable parameter. This method significantly improved the performance of LLaMA-1 (Zhang et al., 2022) and LLaMA-2 (Touvron et al., 2023) on structured fine-tuning tasks.

4 Methods

Unlike traditional RNNs, self-attentions enable models to focus on key interactions within nearby events (local attention) and across temporally distant sets (global attention) through direct connections between all events in a sequence, leading to superior performance.

We start by defining ideal local and global attention mechanisms and then explore how the GLA method enhances robust modeling in complex event sequences.

4.1 Local and Global Attention

Local attention, denoted as Atten_{L_i}, is defined as:

$$\text{Atten}_{L_i} = \bigcup_{j \in N_i} \{e_j\} \quad \text{where} \quad j \geq 1 \tag{5}$$

Here, $\bigcup_{j \in N_i} \{e_j\}$ refers to all sets of events in the neighborhood N_i of a specific event i that exhibit significant interaction within a specific set of input events.

Similarly, global attention, denoted as $\text{Atten}_{G_{i,m}}$, is the intersection between local attentions:

$$\text{Atten}_{G_{i,m}} = \bigcup_{j \in N_i} \{e_j\} \bigcap \bigcup_{n \in N_m} \{e_n\} \quad \text{where} \quad j, n \geq 1 \tag{6}$$

Finally, a global-local self-attention Atten_{GL} is the union of both global and local interactions:

$$\text{Atten}_{GL} = \bigcup \left(\text{Atten}_{G_{i,m}}, \text{Atten}_{L_i}\right) \tag{7}$$

Thus, local attention targets interactions within specific event neighborhoods, global attention captures broader interactions across neighborhoods, and global-local self-attention combines both to provide a comprehensive view of sequence interactions.

4.2 GLA

In this subsection, we introduce a simple two-step augmentation technique, termed Global-Local self-Attention (GLA) (Algorithm 1), designed to enhance the global-local self-attention mechanism within transformer models without altering the computational graph gradient in neural networks, making it applicable to any pre-trained or transformer encoder network:

$$\text{GLA} = \left((A_h + \sim \mathcal{N}(\mu, \sigma_{GN}^2)) * n_{\sigma_{eh}}\right) V \tag{8}$$

$$\mu = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} A_{i,j}}{n^2} \tag{9}$$

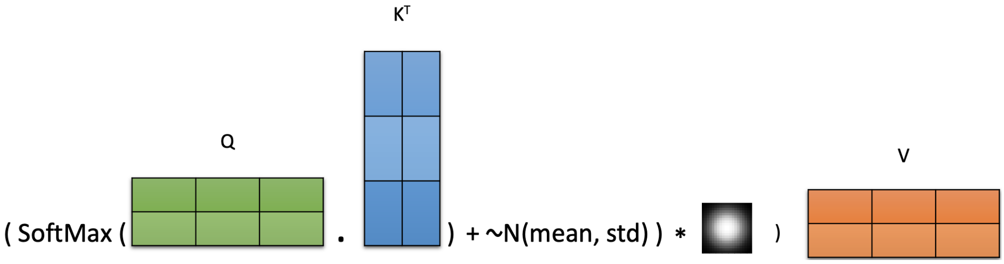


Figure 2: Global-Local self-Attention (GLA) mechanism

$$\sigma_{GN} = \sqrt{\frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(A_{i,j}-\mu)^2}{n-1}} \tag{10}$$

Here, $\sim \mathcal{N}(\mu, \sigma_{GN}^2) \in \mathbb{R}^{n\times n}$ introduces an adaptive normal noise based on the current values of the attention scores. $n_{\sigma_{eh}}[i,j]$ denotes a two-dimensional Gaussian kernel, determining the effective event horizon of neighborhoods (locality), with σ_{eh} being tunable based on the input data's characteristics as:

$$n_{\sigma_{eh}}[i,j] = \frac{1}{2\pi\sigma_{eh}^2} e^{-\frac{1}{2}\left(\frac{i^2+j^2}{\sigma_{eh}^2}\right)} \tag{11}$$

The convolution operation $*$ applies the Gaussian filter to the noised added attention scores, considering the effect of all events in a neighborhood based on their distance as:

$$f[i,j] * n_\sigma[i,j] = \frac{1}{2\pi\sigma^2}\sum_{m=1}^{k} e^{-\frac{1}{2}\left(\frac{m^2}{\sigma^2}\right)} \times \sum_{n=1}^{k} e^{-\frac{1}{2}\left(\frac{n^2}{\sigma^2}\right)} f[i-m, j-n] \tag{12}$$

Furthermore, $k = 2\pi\sigma$ is the kernel size in the convolving operation.

The selection of μ and σ_{GN} is guided by the observation that attention scores better capture temporal changes during fine-tuning. Additionally, since different behaviors are typically observed within each self-attention head, μ and σ_{GN} are calculated and applied separately for each attention head. Figure 2 shows a schematic representation of the GLA mechanism.

Adding normal noise $\sim \mathcal{N}(\mu, \sigma_{GN}^2)$ helps to prevent the model from getting stuck in sub-optimal weights and allows for learning interactions between temporally distant events. Moreover, convolving the result with a Gaussian filter encourages the model to consider the effects of all events in a neighborhood based on their distance.

To ensure that the output of each self-attention head is scaled appropriately and not affected by uncertainty in the normal noise during inference, we deactivate the normal noise and replace it with μ:

$$\text{GLA} = ((A_h + \mu) * n_{\sigma_{eh}})V \tag{13}$$

The computational complexity of GLA is $O(n^2)$ (for more details, see the technical appendix), and since it's primarily used during fine-tuning with limited labeled samples, the additional cost is negligible.

Algorithm 1 A Transformer Encoder with GLA Augmentation

Input: $D_{\text{fine-tuning}} = \{(X_i, y_i)\}_1^N$ tokenized dataset, embedding layer emb$(\cdot)$, attention score matrix A_h, normal noise $\mathcal{N}(\mu, \sigma^2_{\text{GN}})$, two-dimensional Gaussian noise $n_{\sigma_{\text{eh}}}$, rest of the model $f(\cdot)$
Parameter: Normal noise $\mu, \sigma^2_{\text{GN}}$ calculated from A_h, event horizon hyperparameter σ_{eh} based on the neighborhood radius

```
Initialize $\theta$ from a pre-trained model
repeat
    Sample $(X_i, y_i) \sim D_{\text{fine-tuning}}$
    $X_{\text{emb}} \leftarrow \text{emb}(X_i)$
    for each Attention Head $A_h$ in Transformer Block do
        $A_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{emb}}) + \mathcal{N}(\mu, \sigma^2_{\text{GN}})$
        $A_h(X_{\text{attn}}) \leftarrow \text{Convolve}(A_h(X_{\text{attn}}), n_{\sigma_{\text{eh}}})$
        $H_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{attn}})V$
    end for
    $\text{MultiHead}(H) \leftarrow \text{concat}(H_0(X_{\text{attn}}), \dots, H_h(X_{\text{attn}}))$
    $\hat{y}_i \leftarrow f(\text{MultiHead}(H))$
    $\theta \leftarrow \text{opt}(\theta, \text{loss}(\hat{y}_i, y_i))$
until Stopping criteria met or maximum iterations reached
```

5 Experiments

5.1 Datasets

In our study, we utilized medical data from two sources: the MIMIC-IV Johnson et al. (2020) hosp module and the Malmö Diet and Cancer Cohort (MDC) Berglund et al. (1993) dataset, approved by the Ethics Review Board of Sweden (Dnr 2023-00503-01). Each EHR trajectory represents a sequence of temporally structured health events. The MIMIC-IV dataset includes 173,000 patient records across 407,000 visits from 2008 to 2019, with 10.6 million medical codes. The MDC dataset, from a cohort study in Sweden, comprises 30,000 individuals with 531,000 visits from 1992 to 2020, offering a more extended patient history—257 codes per patient on average, compared to MIMIC-IV's 61. To ensure consistency, we used only ICD and ATC codes, the only types available in MDC at the beginning, aligning with prior work like Med-BERT on diagnosis codes for risk prediction.

Both datasets use ICD and ATC codes for disease and medication classification. We randomly split each cohort into 70% for pre-training, 20% for fine-tuning, and 10% for testing. After preprocessing, MIMIC-IV had 2,195 unique ICD-9 and 137 ATC-5 codes, while MDC had 1,558 ICD-10 and 111 ATC-5 codes. To assess the generalizability and robustness of our results, the fine-tuning dataset was split into 5 folds. The model was fine-tuned on 4 folds with early stopping on the remaining fold, repeated 5 times with different validation sets. We reported the mean and standard deviation of the AUC on the unseen test dataset. For details, refer to the dataset specifications and implementation details in the technical appendix.

5.2 Problem Formulation

Each dataset D comprises a set of patients P, $D = \{P^1, P^2, \dots, P^{|D|}\}$. In our study, we considered a total of $|D| = 172,980$ patients for MIMIC-IV and $|D| = 29,664$ patients for the MDC cohort. We represent each patient's longitudinal medical trajectory through a structured set of visit encounters as a sequence of events. This representation is denoted as $P^i = \{V_1^i, V_2^i, \dots, V_O^i\}$, where O represents the total number of visit encounters for patient i. Each visit $V_j^i = I_j \cup M_j$ is the union of all diagnosis codes $I_j \subset I$ and prescribed medications $M_j \subset M$ that are recorded for the P^i at visit V_j^i. To reduce sparsity, we excluded less frequently occurring medical codes and retained only the initial 4 digits of ICD and ATC codes.

To guide the model in understanding changes in encounter times and the structure of each patient's trajectory, similar to BERT, we employed special tokens. A $[CLS]$ token is placed at the beginning of each patient's trajectory, while a $[SEP]$ token is inserted between visits. Consequently, each patient's trajectory is represented as $P^i = \{[CLS], V_1^i, [SEP], V_2^i, [SEP], \ldots, V_O^i, [SEP]\}$, providing the model with valuable context for analysis and prediction.

Here, we evaluated our models on 3 downstream tasks e_{dt} (Heart Failure (HF), Alzheimer Disease (AD), Prolonged Length of Stay on the next visit (PLS) predictions), where the model predicts the incidence of the first HF ($I_{N=HF}$) or AD ($I_{N=AD}$) ICD codes or the presence of PLS ($PLS_N = 1$) on the N^{th} visit, given the patient's previous history of medical codes, $[V_1^i : V_{N-1}^i]$, as a sequence of temporally structured health events:

$$\mathbb{P}(e_{dt} \in V^N \mid P^i = \{[CLS], V_1^i, [SEP], V_2^i, [SEP], \ldots, V_{N-1}^i, [SEP]\}) \tag{14}$$

For each patient's trajectory, if there were no occurrences of the target events e_{dt}, it is considered a negative case; otherwise, we exclude the first visit with the target and all subsequent visits and consider it a positive case. All ATC codes related to HF treatment are excluded to avoid timing-related noise and non-trivial predictions. Initially, models exhibited bias toward longer visit histories, confounding risk predictions. To address this, we excluded trajectories with fewer than 30 visits in the MDC dataset and fewer than 10 visits in the MIMIC-IV dataset. This ensured balanced visit histories between positive and negative cases, resulting in averages of 19 visits in the MDC dataset and 9 visits in the MIMIC-IV dataset, aligning with their overall dataset averages prior to preprocessing. Table 1 summarizes the number of positive and negative cases after these preprocessing steps.

Table 1: Number of positive and negative labeled samples in each downstream task.

Task	**#Positive labels**	**#Negative labels**
PLS prediction	2429	6360
HF prediction on the MIMIC IV	243	641
AD prediction	245	2628
HF prediction on the MDC	103	301

5.3 List of Models

To thoroughly investigate the impact of the proposed GLA regularization, we compared the performance of following conventional and deep learning models on downstream tasks of HF, AD, and PLS prediction using both the MDC and MIMIC-IV datasets. These models were trained either from scratch or initiated from pre-trained weights, fine-tuned on the fine-tuning dataset, and evaluated on the test dataset. We set the tunable event horizon parameter to $\sigma_{eh} = 1.0$ (kernel size = 7) for the GLA regularization on the MDC dataset and $\sigma_{eh} = 0.33$ (kernel size = 3) on the MIMIC IV after fine-tuning on the fine-tuning dataset.

Baseline Models The baselines in our study were selected based on prior research and practical considerations for modeling temporal health data. The following models were used for baseline comparison: Logistic regression (LR), random forest (RF), multilayer perceptron (MLP), bidirectional gated recurrent unit (Bi-GRU), transformer encoder with multi-head attention and a classification feedforward layer at the top, trained from scratch, and finally a transformer encoder pre-trained on MLM followed by fine-tuning of all weights for downstream tasks (Rasmy et al., 2021; Li et al., 2020; Meng et al., 2021). For LR, RF, and MLP we encoded each visit as a multi-hot vector and aggregated visits by summing them. Additionally, MLM pre-trained transformers, similar to BEHRT and Med-BERT, formed the foundation for representing temporal health data, enabling us to compare GLA's performance and demonstrate its ability to enhance existing methods without major structural changes.

Models with proposed GA/GLA

- Transformer with GLA augmentation: This model incorporates GLA into all self-attention heads of the randomly initialized transformer.
- Transformer pre-trained on MLM with Global Attention (GA): In this approach, $\mathcal{N}(\mu, \sigma_{GN}^2)$ (Global term) is added to all self-attention heads of a pre-trained transformer. This experiment allows us to isolate the impact of the normal noise and the Gaussian kernel convolution operations.
- Transformer pre-trained on MLM with GLA : This model incorporates GLA into all self-attention heads of the pre-trained transformer.

5.4 Evaluation on downstream tasks

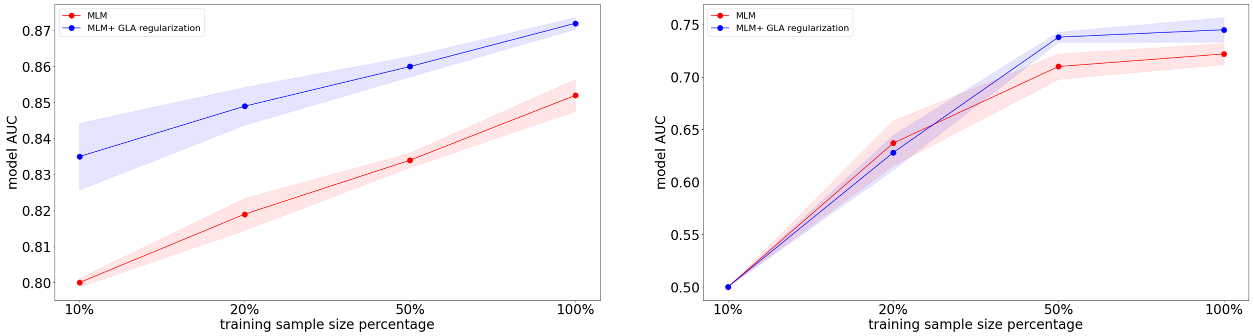
The results are summarized in Table 2 and suggests that adding GLA improves the AUC of pre-trained transformers, potentially positioning them as one of the state-of-the-art methods for outcome prediction on temporal structured health data. Specifically, on the MDC dataset, the AUC for HF and AD prediction increased to 74.5% and 73.2%, respectively, while on the MIMIC-IV dataset, the AUC for HF prediction reached 87.2%. The addition of GLA resulted in statistically significant improvements for HF prediction on both the MDC and MIMIC-IV datasets for the MLM pre-trained transformer. Furthermore, the improvement in AD prediction was considerable, showcasing the effectiveness of GLA augmentation. However, incorporating GLA did not significantly alter the performance of PLS prediction. Additionally, applying GLA to randomly initialized transformers boosted the AUC for PLS prediction to 60.2%, with negligible effects on other downstream tasks. To delve deeper into the impact of each local and global augmentation term – Gaussian kernel and normal noise, respectively – we solely added the normal noise to the pre-trained transformer. This experiment revealed that the global term alone had a more pronounced effect on downstream tasks in the MIMIC dataset, whereas the combined (GLA) terms exhibited greater impacts on the downstream tasks in the MDC dataset, particularly associated with its longer sequences. This suggests that emphasizing locality through GLA is especially beneficial for handling longer sequences.

Table 2: Average AUC values (%) and standard deviation for different methods for the HF prediction, AD prediction, and PLS prediction downstream tasks on the test datasets.

Model / Downstream Task	HF prediction (MDC)	AD prediction (MDC)	HF prediction (MIMIC-IV)	PLS prediction (MIMIC-IV)
Logistic regression	62.4 (1.1)	56.4 (1.1)	83.9 (1.2)	54.2 (0.4)
Random forest	60.7 (0.5)	51.8 (0.3)	77.2 (2.3)	51.1 (0.3)
MLP	67.9 (3.0)	68.0 (1.5)	85.2 (0.3)	59.3 (1.9)
Bi-GRU	62.3 (1.2)	60.4 (1.1)	86.5 (1.2)	55.9 (1.0)
Transformer	71.4 (0.5)	70.5 (0.8)	84.2 (1.4)	54.4 (0.8)
Transformer+ GLA	72.1 (2.7)	70.4 (0.6)	83.2 (2.5)	60.2 (1.2)
Transformer pre-trained on MLM	72.2 (2.5)	72.2 (1.1)	85.2 (1.1)	60.3 (1.3)
transformer pre-trained on MLM+ GA	72.6 (1.9)	71.4 (1.0)	86.5 (1.2)	**60.7 (0.6)**
transformer pre-trained on MLM+ GLA	**74.5 (2.9)**	**73.2 (0.3)**	**87.2 (0.4)**	60.3 (0.7)

5.5 Performance boost on data insufficiency

One of the advantages of using pre-trained transformers is their robustness and performance in situations of data insufficiency, observed in both NLP (Brown et al., 2020) and temporal health data (Rasmy et al., 2021). Here, we investigated the effect of applying GLA on model performance for HF prediction with reduced data sample sizes. We decreased the fine-tuning sample size to 50%, 20%, and 10%, respectively. The performance of the pre-trained transformer with and without GLA, was compared on both the MDC and MIMIC-IV datasets. Figure 3a shows that GLA improves the model performance by around 3% in HF



(a) AUC values for HF prediction across various fine-tuning sample sizes on the test dataset in MIMIC IV.

(b) AUC values for HF prediction across various fine-tuning sample sizes on the test dataset in the MDC.

Figure 3: GLA's impact on HF prediction across fine-tuning sample sizes in MIMIC IV and MDC datasets.

prediction on the MIMIC-IV dataset across all data sample sizes. Similarly, Figure 3b demonstrates that GLA consistently outperforms the baseline in HF prediction on the MDC dataset, even with a 50% reduction in training samples. However, its superiority diminishes with less data.

5.6 VS Naive masking

Randomly masking the attention score matrix during training can be seen as an extreme form of GA augmentation. Instead of adding normal noise to perturb relationships between events in a sequence, naive masking directly disrupts these relationships by summing each element with 0 or $-A_{h_{i,j}}$, effectively breaking the connections between tokens. We compared our method with naive self-attention masking, as described by Wu et al. (2023), which introduces a bias in the structure of self-attentions:

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}} + M\right), \quad M \in \{0, -\infty\}^{N \times N}, \tag{15}$$

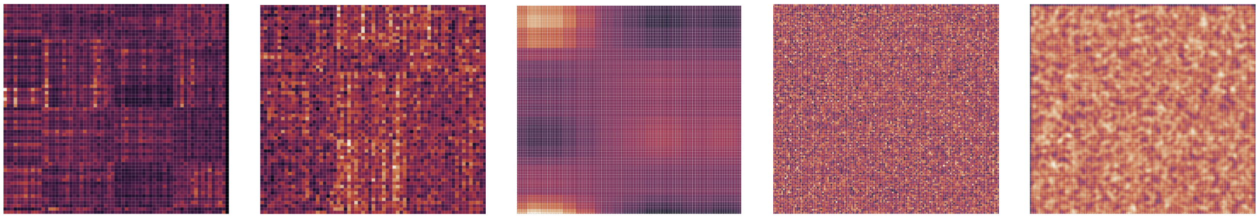
where $M_{i,j} = -\infty$ with $p = 0.2$, optimized based on performance on the fine-tuning dataset. We extended it to DropAttention (Zehui et al., 2019), which expands the mask with a span length ω and we set $\omega =$ Kernel size. However, neither naive masking nor DropAttention improved the performance of the pre-trained transformer for HF prediction on the MDC and MIMIC-IV datasets. Instead, these methods only increased the number of training iterations required for convergence (see Table 3). While these techniques can help mitigate overfitting, their overly aggressive regularization often disrupts critical dependencies within sequences, leading to unstable training and poorer overall performance, especially on complex healthcare prediction tasks. In contrast, GLA provides controlled regularization that balances the attention distribution and prevents over-reliance on specific patterns, preserving essential relationships within the data and promoting more effective representations(see the appendix for a detailed justification for GLA).

5.7 VS Noisy embedding augmentation

We also compared GLA with other noise augmentation methods, specifically evaluating the impact of adding noise to different layers of the transformer, including the embedding layer (as done in NefTune Jain et al. (2023)). Additionally, we explored noise augmentation in the feedforward layer and compared these approaches to GLA. As shown in Table 3, although NefTune enhances the performance of pre-trained transformers in HF prediction across both datasets, GLA consistently outperforms both NefTune and feedforward noise augmentation in predicting outcomes. While GLA demonstrates superior performance in this context, NefTune has the advantage of being computationally lighter. However, since both methods are applied during fine-tuning, the computational demands are not a significant concern.

Table 3: Comparing GLA with naive masking and augmentation methods. The table shows the average AUC values (%) and standard deviation across HF prediction tasks on the MDC and MIMIC-IV datasets.

Model / Downstream Task	HF Prediction (MDC)	HF Prediction (MIMIC-IV)
Transformer pre-trained on MLM	72.2 (2.5)	85.2 (1.1)
Transformer pre-trained on MLM+ Naive masking	70.00 (1.5)	85.1 (0.7)
Transformer pre-trained on MLM+ DropAttention	69.7 (1.1)	84.9 (1.3)
Transformer pre-trained on MLM+ NEFTune($\alpha = 5$)	73.6 (3.2)	85.2 (0.7)
Transformer pre-trained on MLM+ NEFTune($\alpha = 10$)	73.1 (1.7)	85.5 (0.4)
Transformer pre-trained on MLM+ noise in the feedforward($\alpha = 5$)	73.7 (2.2)	85.0 (1.2)
Transformer pre-trained on MLM+ noise in the feedforward($\alpha = 10$)	72.5 (4.4)	84.5 (0.8)
Transformer pre-trained on MLM+ GLA regularization	**74.5 (2.9)**	**87.2 (0.4)**



(a) Transformer (b) Transformer+GLA (c) Pre-trained Transformer (d) Pre-trained Transformer+GA (e) Pre-trained Transformer+GLA

Figure 4: Comparing the impact of GLA on the self-attention score weights for five fine-tuned models on HF prediction on the MDC dataset for a specific test sample. The scale of the heatmaps varies across different models.

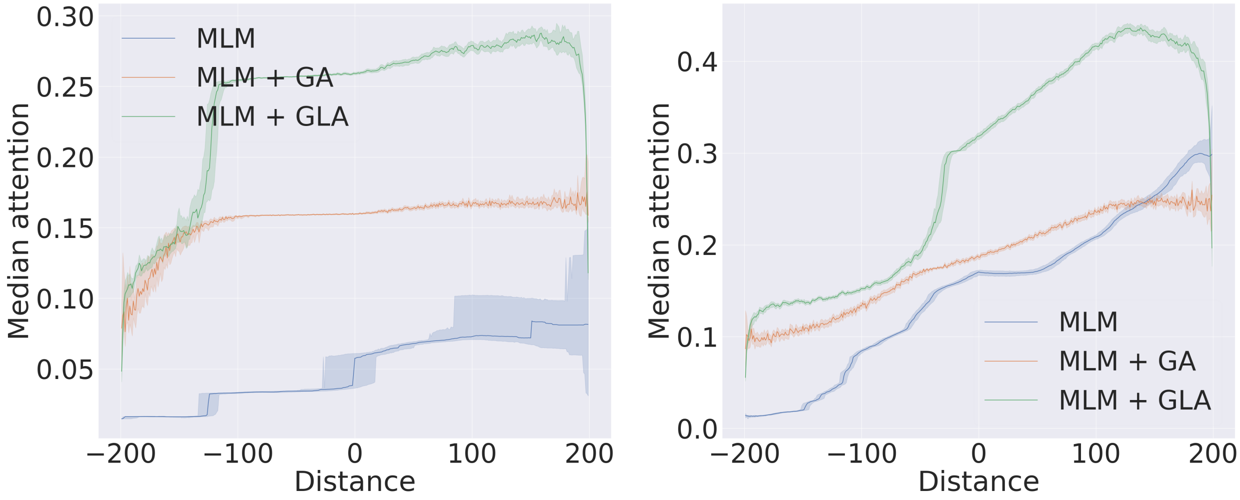
5.8 Effect of GLA on self-attention behavior

Analyzing self-attention weights and attention score matrices can highlight how transformers prioritize relationships between events, shedding light on their internal logic and behavior (Clark et al., 2019; Kovaleva et al., 2019; Hao et al., 2021). To explore the impact of GLA on the transformers, we visualized the attention score matrices A_h and $A_{h-GLA} = (A_h + \mu) * n_{\sigma_{eh}}$ of all transformer-based model for a specific sample from the test dataset fine-tuned for predicting HF on the MDC datasets (see appendix for effect on the MIMIC-IV dataset).

Light dots in the upper and lower right corners of the attention matrix indicate instances of global attention, where early events assign more weight to temporally distant events. Figure 4 shows the effect of augmenting pre-trained transformers with GA and GLA. GA, alone increases global attention, while it often results in unstructured noise. GLA, however, allows the model to attend adequately to both near and far events, creating a more balanced and structured model. Additionally, incorporating GLA into a randomly initialized transformer leads to the emergence of richer structural patterns.

5.8.1 Effect of GLA on the Receptive Field

The self-attention mechanism is designed to capture both long and short-range dependencies effectively. To quantitatively assess the impact of GA and GLA on the receptive field, we plot the median values of attention score matrix A_h for each event with respect to all previous and subsequent events $(i - j, A_{h_{i,j}})$ -i, j are positions of e_i, e_j in the sequence of events- across all test samples for HF and AD predictions



HF prediction

AD prediction

Figure 5: Impact of GLA on the receptive field of the self-attentions for HF and AD prediction on the MDC dataset.

on the MDC (Figures 5). Transformers pre-trained on MLM typically allocate more attention weight to recent events, often in a monotonous fashion. Incorporating GA regularization reduces the steepness of this attention distribution, allowing events to receive more balanced attention, not solely based on their proximity to recent events. Ultimately, applying GLA regularization preserves the benefits of GA by providing a more equal distribution of attention within a local neighborhood, while simultaneously reducing the emphasis on very distant past events. However, it is important to note that raw self-attention values do not fully reveal transformer behavior, as they are not directly interpretable and require further processing for accurate attribution (Hao et al., 2021; Jain & Wallace, 2019; Serrano & Smith, 2019).

6 Conclusion

Transformers' ability to model both nearby and distant interactions within a sequence enhances performance on complex data with limited samples. However, the complexity of interactions combined with small sample sizes can sometimes degrade the effectiveness of pre-trained transformers. We introduced the GLA augmentation method, which seamlessly integrates into any pre-trained transformer during fine-tuning without altering the computational graph. GLA enhances self-attention by capturing both local and global sequence complexities through adaptive noise injection and Gaussian kernel smoothing. Our results show that GLA consistently boosts performance on downstream tasks, enhances robustness with limited data, and better balances the attention distribution. While this work focused on healthcare applications, GLA's flexibility and domain-agnostic design make it applicable across various transformer-based models. By bridging performance and practicality, GLA offers a scalable solution to enhance transformer performance within complex data, paving the way for further exploration across diverse fields.

References

Ali Amirahmadi, Mattias Ohlsson, and Kobra Etminani. Deep learning prediction models based on ehr trajectories: A systematic review. *Journal of biomedical informatics*, pp. 104430, 2023.

Ali Amirahmadi, Farzaneh Etminani, Jonas Bjork, Olle Melander, and Mattias Ohlsson. Too-bert: A trajectory order objective bert for self-supervised representation learning of temporal healthcare data. 2024a.

Ali Amirahmadi, Mattias Ohlsson, Kobra Etminani, Olle Melander, and Jonas Björk. A masked language model for multi-source ehr trajectories contextual representation learning. *arXiv preprint arXiv:2402.06675*, 2024b.

G Berglund, S Elmståhl, L Janzon, and SA Larsson. The malmo diet and cancer study. design and feasibility. *Journal of internal medicine*, 233(1):45–51, 1993.

Heloísa Oss Boll, Ali Amirahmadi, Mirfarid Musavian Ghazani, Wagner Ourique de Morais, Edison Pignaton de Freitas, Amira Soliman, Kobra Etminani, Stefan Byttner, and Mariana Recamonde-Mendoza. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Journal of Biomedical Informatics*, pp. 104616, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 606–613, 2020.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, 2023.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12963–12971, 2021.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*, 2023.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

HyunGi Kim, Seungryong Yoo, Bong Gyun Kang, Saehyung Lee, and Sungroh Yoon. Protoreg: Prioritizing discriminative information for fine-grained transfer learning. 2023.

Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin Taylor, and Tom Goldstein. Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 60–69, 2022.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, Zeyu Chen, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. *arXiv preprint arXiv:1901.09229*, 2019.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25(8):3121–3129, 2021.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pp. 239–260. PMLR, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Houxing Ren, Jingyuan Wang, Wayne Xin Zhao, and Ning Wu. Rapt: Pre-training of time-aware transformer for learning robust healthcare representation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3503–3511, 2021.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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, 2023.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. Adversarial self-attention for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13727–13735, 2023.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pp. 2825–2834. PMLR, 2018.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4133–4141, 2017.

Kaichao You, Zhi Kou, Mingsheng Long, and Jianmin Wang. Co-tuning for transfer learning. *Advances in Neural Information Processing Systems*, 33:17236–17246, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Lin Zehui, Pengfei Liu, Luyao Huang, Junkun Chen, Xipeng Qiu, and Xuanjing Huang. Dropattention: A regularization method for fully-connected self-attention networks. *arXiv preprint arXiv:1907.11065*, 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Nan Zhou, Jiaxin Chen, and Di Huang. Dr-tune: Improving fine-tuning of pretrained visual models by distribution regularization with semantic calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1547–1556, 2023.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

Weicheng Zhu and Narges Razavian. Variationally regularized graph-based representation learning for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 1–13, 2021.

A Appendix

A.1 Implementation details

The code compatible with the public MIMIC-IV dataset is available in the Code Appendix.

We initialized our model with a pre-trained transformer encoder block with 5 heads on the MLM task. For fine-tuning, we aggregated token representations with a GRU layer and fed them into the classifier. The Adam optimizer with layer-wise learning rate decay was used (coef=0.9, initial learning rate=6e-5). The input length was set to 200 medical codes. Details of $\sigma_e h$ and kernel size are described in the Method Section. Cross-Validation: The fine-tuning dataset was split into 5 folds. The model was fine-tuned on 4 folds with early stopping on the remaining fold, repeated 5 times with different validation sets. We reported the mean and std of the AUC on the unseen test dataset.

A.2 Dataset specifications

We used medical data from two sources: the Medical Information Mart for Intensive Care IV (MIMIC-IV) Johnson et al. (2020) hosp module, and the Malmö Diet and Cancer Cohort (MDC) Berglund et al. (1993) dataset, approved by the Ethics Review Board of Sweden (Dnr 2023-00503-01). Each EHR trajectory represents a sequence of events of temporal structured health data. The MIMIC-IV hosp module is a comprehensive collection of inpatient EHR trajectories, containing approximately 173,000 patient records documented during 407,000 visits spanning from 2008 to 2019. This dataset includes a total of 10.6 million medical codes representing diagnoses and medications.

On the other hand, the MDC dataset originates from a prospective cohort study conducted in Sweden. It consists of around 30,000 individuals residing in Malmö between 1991 and 1996, with records of both inpatient and outpatient visits spanning from 1992 to 2020, resulting in a total of 531,000 visits. While the MDC dataset has fewer overall samples, it provides a more extensive patient history, with an average of 257 codes per patient compared to MIMIC-IV's 61.

Both datasets use the International Statistical Classification of Diseases and Related Health Problems (ICD) and Anatomical Therapeutic Chemical Code (ATC) for disease and medication classification, respectively, in a hierarchical format.

To facilitate our self-supervised pre-training, supervised fine-tuning, and final testing, we partitioned the extracted cohort randomly into three subsets: 70%, 20%, and 10%, respectively. Despite being characterized by extensive sparsity, preprocessing resulted in 2,195 unique ICD-9 and 137 unique ATC-5 codes for the MIMIC-IV dataset and 1,558 unique ICD10 and 111 unique ATC-5 codes for the MDC dataset.

Table 4: MIMIC-IV dataset summary statistics.

	Pre-training dataset	Fine-tuning dataset	Test dataset	Total dataset
#patients	121 K	36 K	16 K	173 K
#visits	285 K	86 K	37 K	408 K
#Medical codes	7.451 M	2.234 M	937 K	10.622 M

Table 5: MDC dataset summary statistics.

	Pre-raining dataset	Fine-tuning dataset	Test dataset	Total dataset
#patients	21 K	6 K	3 K	30 K
#visits	373 K	107 K	52 K	531 K
#Medical codes	5.339 M	1.554 K	741 K	7.634 M

A.2.1 Data availability

The MIMIC-IV data is available on `https://physionet.org/content/mimiciv/2.2/`. The MDC dataset is available upon application and with permission of the Malmo Population-Based Cohorts Joint Database `https://www.malmo-kohorter.lu.se/malmo-cohorts`

A.3 Justification for GLA: A Comparison with Dropout

In addition to its demonstrated global-local impact, GLA can be justified by drawing parallels with dropout regularization, viewing GLA as an "adaptive" extension of it. Dropout works by randomly setting some of the activations to zero, effectively disconnecting certain nodes during training. This prevents the model from becoming overly dependent on specific neurons and encourages a more robust and generalized representation.

Mathematically, for each attention head A_h, dropout can be seen as applying a mask M (where M is a Bernoulli distribution), resulting in the modified attention head $A'_h = M \cdot A_h$. In contrast, GLA applies a more nuanced adjustment:

$$A_h = A_h + \epsilon \sim \mathcal{N}(\mu, \sigma^2) = \left(1 + \frac{\epsilon}{A_h}\right) A_h = P \cdot A_h \tag{16}$$

Here, $P = \left(1 + \frac{\epsilon}{A_h}\right)$ acts as an adaptive perturbation factor. Instead of completely severing connections between tokens (as in dropout), GLA adjusts the attention weights by either amplifying or diminishing the focus between two events. This approach maintains the relationships within the data while still introducing variability.

The random perturbation from GLA forces the attention mechanism to avoid over-reliance on specific patterns by continually adjusting the attention distribution. Consequently, GLA can be seen as a form of ensemble learning, where each perturbation offers a different perspective on the data. This effectively trains multiple versions of the model in parallel, each slightly varied due to the noise, leading to a more robust and generalized final model.

A.4 Performance boost on data insufficiency

Table 6 presents the numerical results corresponding to the data insufficiency section.

Table 6: Effect of incorporating GLA into the pre-trained transformer on AUC performance value of HF prediction across various fine-tunning sample sizes on the test dataset in MIMICIV and the MDC

Model-dataset / fine-tuning percentage	10 %	25%	50%	100%
MLM-MDC	0.5 (0)	0.637 (0.053)	0.710 (0.030)	0.722 (0.025)
MLM+GLA-MDC	0.5 (0)	0.628 (0.041)	0.738 (0.012)	0.745 (0.029)
MLM-MIMICIV	0.800 (.003)	0.819 (.011)	0.834 (.005)	0.852 (.011)
MLM+GLA-MIMICIV	0.835 (.023)	0.849 (.013)	0.860 (.007)	0.872 (.004)

A.5 Computational Complexity and Scalability

GLA involves two primary operations per attention head: adding iid normal noise to an $n \times n$ attention score matrix and convolving the result with a Gaussian filter. Therefore, the computational complexity of GLA can be expressed as:

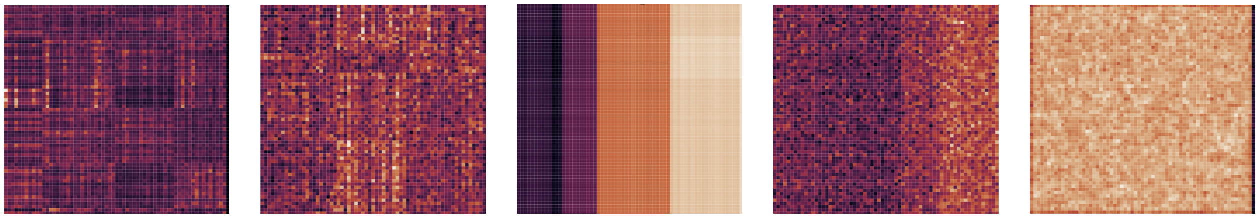
$$O(\text{GLA}) = O(\text{addition of } n \times n \text{ matrix}) + O(\text{convolution})$$

$$O(\text{GLA}) = O(n^2) + O(k^2 \times n^2) = O(n^2) + O(n^2) = O(n^2)$$

Here, n represents the input length (where $n = 200$ in our case), and k is the kernel size ($k \ll n$). Since the noise addition and convolution operations are only performed during the fine-tuning phase—where the number of samples is significantly smaller compared to pre-training—GLA introduces minimal scalability limitations.

A.6 Effect of GLA on self-attention behavior on the MIMIC-IV dataset

Figure 6 shows the effect of augmenting pre-trained transformers with GA and GLA on a specific sample on the MIMIC-IV dataset.



(a) Transformer (b) Transformer+GLA (c) Pre-trained Transformer (d) Pre-trained Transformer+GA (e) Pre-trained Transformer+GLA

Figure 6: The attention score weights for ten fine-tuned models on HF prediction on the MIMIC-IV dataset for a specific sample.scale of the heatmaps varies across different models.