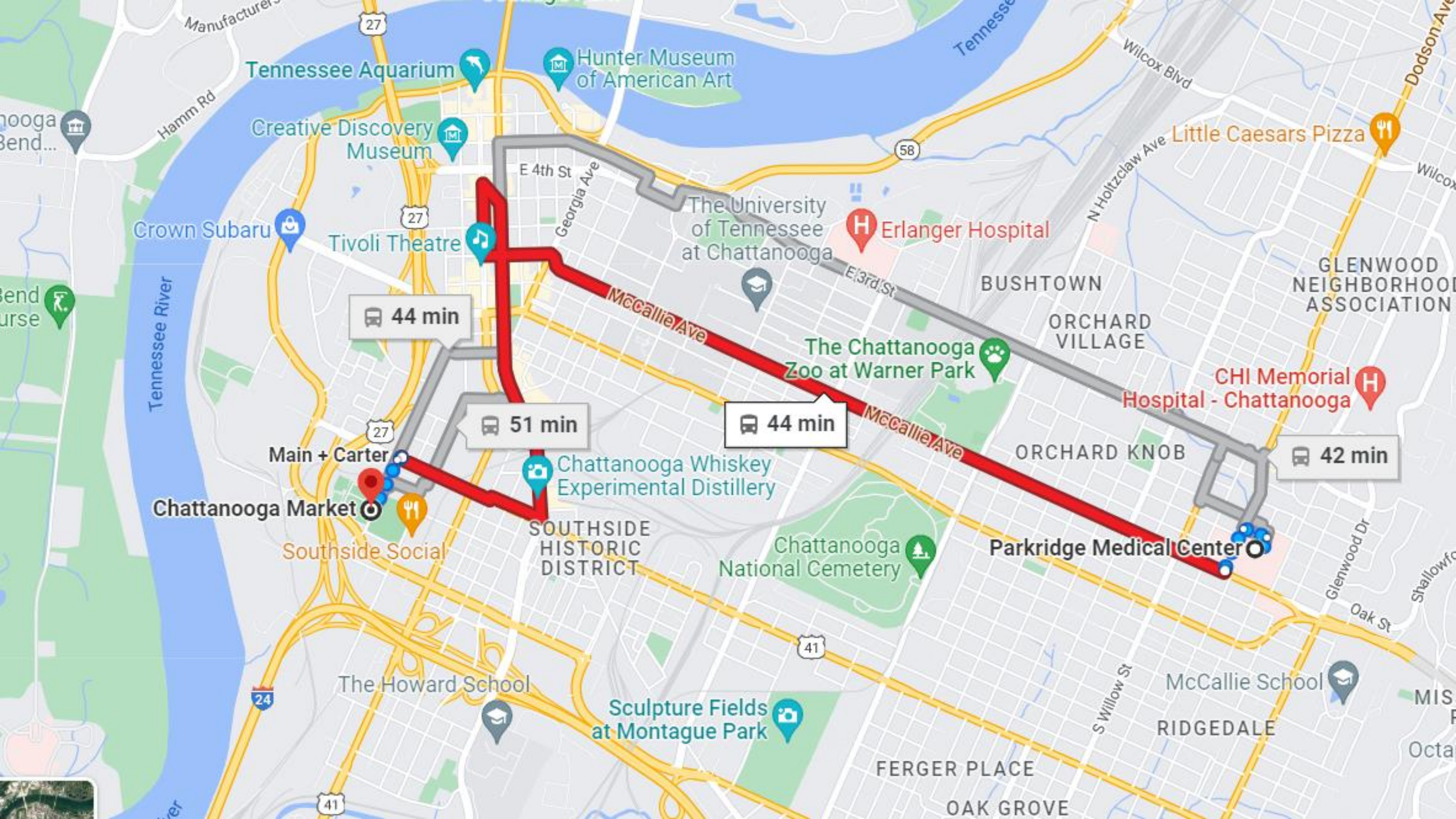
Transit Delay Time prediction using Time Series Transformers

Pravesh Koirala



Tennessee Aquarium
Hunter Museum of American Art
Creative Discovery Museum
E 4th St
Georgia Ave
The University of Tennessee at Chattanooga
Erlanger Hospital
Little Caesars Pizza
Wilcox Blvd
N Holtzclaw Ave
Dodson Ave
Crown Subaru
Tivoli Theatre
Tennessee River
Hamm Rd
44 min
Mccallie Ave
E 3rd St
BUSHTOWN
GLENWOOD NEIGHBORHOOD ASSOCIATION
ORCHARD VILLAGE
The Chattanooga Zoo at Warner Park
CHI Memorial Hospital - Chattanooga
44 min
51 min
42 min
ORCHARD KNOB
Main + Carter
Chattanooga Market
Chattanooga Whiskey Experimental Distillery
SOUTHSIDE HISTORIC DISTRICT
Southside Social
Parkridge Medical Center
Chattanooga National Cemetery
Glenwood Dr
Oak St
The Howard School
Sculpture Fields at Montague Park
S Willow St
McCallie School
RIDGEDALE
FERGER PLACE
OAK GROVE
27
58
41
24

GIONAL
ORITY

About | Routes/Schedules | Fares | Using CARTA | Contact

CHATTANOOGA'S
RIDE

Problem Statement

- Predict the delay in the Fixed-line Transit System.
- Predictors?
  - Weather (temperature, humidity, precipitation)
  - Day of the week.
  - Time of the day.
  - School breaks / holidays / Concerts or Events.
  - Nature of the route.
    - Downtown vs Suburb.

Challenges

- Nature of Predictors:

Predictor	Cardinality
Temperature	$\mathbb{R}^+$
Humidity	$\mathbb{R}^+$
Precipitation	$\mathbb{R}^+$
Travel Time	$\mathbb{R}^+$
Segment Length	$\mathbb{R}^+$
School Break	$\{0, 1\}$
National Holiday	$\{0,1\}$
Day of the Week	$\{Mon, Tue, \dots Sun\}$
Time of the Day (15-min bucket)	$\{0, 1, 2, \dots 96\}$
Route Segment	$\{S_1, S_2, \dots S_{50}\}$

**There are more than 3000 route segments in the dataset, we only consider a few for the purpose of simplification*

Time Series Transformer

- An extension of the *traditional* transformer that allows you to work with continuous time predictors.
- Check https://github.com/PraveshKoirala/Transformers-Paper for an overview on the method.
- **EXTENSION:**
  - Change the architecture in such a way that it now works with both Categorical and Continuous predictors.

Algorithm for Time Series Forecasting

Input: $\mathbf{z}, \mathbf{x} \in \mathbb{R}^*$, two sequences of IL-Ratio inputs (+ve reals).

Output: $\boldsymbol{P} = \mathbb{R}^{length(X)}$ where $p_t \in (0,1)$ is the t-th IL-Ratio.

Hyperparameters: $l_{max} = 10, L_{enc} = 4, L_{dec} = 4, H, d_e = d_{model}, d_{mlp} \in \mathbb{N}$

Parameters: $\boldsymbol{\theta}$ includes all of the following parameters:

$\boldsymbol{W_e} \in \mathbb{R}^{d_e \times 1}, \boldsymbol{W_p} \in \mathbb{R}^{d_e \times l_{max}}$, the token and positional embedding matrices.

For $l \in [L_{enc}]$:

$\boldsymbol{W}_l^{enc}$, multi-head encoder attention parameters for layer l

$\boldsymbol{\gamma}_l^1, \boldsymbol{\beta}_l^1, \boldsymbol{\gamma}_l^2, \boldsymbol{\beta}_l^2 \in \mathbb{R}^{d_e}$, two sets of layer-norm parameters,

$\boldsymbol{W}_{mlp_1}^l \in \mathbb{R}^{d_{mlp} \times d_e}, \boldsymbol{b}_{mlp}^l \in \mathbb{R}^{d_{mlp}}, \boldsymbol{W}_{mlp2}^l \in \mathbb{R}^{d_e \times d_{mlp}}, \boldsymbol{b}_{mlp2}^l \in \mathbb{R}^{d_e}$, MLP parameters.

For $l \in [L_{dec}]$:

$\boldsymbol{W}_l^{dec}$, multi-head decoder attention parameters for layer l

$\boldsymbol{W}_l^{e/d}$, multi-head cross-attention parameters for layer l.

$\boldsymbol{\gamma}_l^3, \boldsymbol{\beta}_l^3, \boldsymbol{\gamma}_l^4, \boldsymbol{\beta}_l^4 \in \mathbb{R}^{d_e}$, two sets of layer-norm parameters,

$\boldsymbol{W}_{mlp_3}^l \in \mathbb{R}^{d_{mlp} \times d_e}, \boldsymbol{b}_{mlp}^l \in \mathbb{R}^{d_{mlp}}, \boldsymbol{W}_{mlp4}^l \in \mathbb{R}^{d_e \times d_{mlp}}, \boldsymbol{b}_{mlp4}^l \in \mathbb{R}^{d_e}$, MLP parameters.

$\boldsymbol{W_u} \in \mathbb{R}^{length(X) \times d_e}$, the unembedding matrix.

/* Encoder portion */

$l_z \leftarrow length(\boldsymbol{z})$

for $t \in [l_z]$: $\boldsymbol{e}_t \leftarrow \boldsymbol{W}_e[:, z[t]] + \boldsymbol{W}_p[:, t]$

$\boldsymbol{Z} \leftarrow [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_{l_z}]$

for $l = 1, 2, \ldots, L_{enc}$ **do**

$\quad \boldsymbol{Z} \leftarrow \boldsymbol{Z} + MHAttention(\boldsymbol{Z} \mid \boldsymbol{W}_l^{enc}, Mask \equiv 1)$

$\quad$ for $t \in [l_z]$: $\boldsymbol{Z}[:, t] \leftarrow layer\_norm(\boldsymbol{Z}[:, t] \mid \boldsymbol{\gamma}_l^1, \boldsymbol{\beta}_l^1)$

$\quad \boldsymbol{Z} \leftarrow \boldsymbol{Z} + \boldsymbol{W}_{mlp2}^l ReLU(\boldsymbol{W}_{mlp1}^l \boldsymbol{Z} + \boldsymbol{b}_{mlp1}^l \boldsymbol{1}^T) + \boldsymbol{b}_{mlp2}^l \boldsymbol{1}^T$

$\quad$ for $t \in [l_z]$: $\boldsymbol{Z}[:, t] \leftarrow layer\_norm(\boldsymbol{Z}[:, t] \mid \boldsymbol{\gamma}_l^2, \boldsymbol{\beta}_l^2)$

end

/* Decoder portion */

$l_x \leftarrow length(\boldsymbol{x})$

for $t \in [l_X]$: $\boldsymbol{e}_t \leftarrow \boldsymbol{W}_e[:, x[t]] + \boldsymbol{W}_p[:, t]$

$\boldsymbol{X} \leftarrow [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_{l_X}]$

for $l = 1, 2, \ldots, L_{dec}$ **do**

$\quad \boldsymbol{X} \leftarrow \boldsymbol{X} + MHAttention(\boldsymbol{X} \mid \boldsymbol{W}_l^{dec}, Mask[t, t'] \equiv [[t \le t']])$

$\quad$ for $t \in [l_X]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] \mid \boldsymbol{\gamma}_l^3, \boldsymbol{\beta}_l^3)$

$\quad \boldsymbol{X} \leftarrow \boldsymbol{X} + MHAttention(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{W}_l^{e/d}, Mask \equiv 1)$

$\quad$ for $t \in [l_z]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] \mid \boldsymbol{\gamma}_l^4, \boldsymbol{\beta}_l^4)$

$\quad \boldsymbol{X} \leftarrow \boldsymbol{X} + \boldsymbol{W}_{mlp4}^l ReLU(\boldsymbol{W}_{mlp3}^l \boldsymbol{Z} + \boldsymbol{b}_{mlp3}^l \boldsymbol{1}^T) + \boldsymbol{b}_{mlp4}^l \boldsymbol{1}^T$

$\quad$ for $t \in [l_X]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] \mid \boldsymbol{\gamma}_l^5, \boldsymbol{\beta}_l^5)$

end

return $\boldsymbol{P} = \boldsymbol{W_u X}$

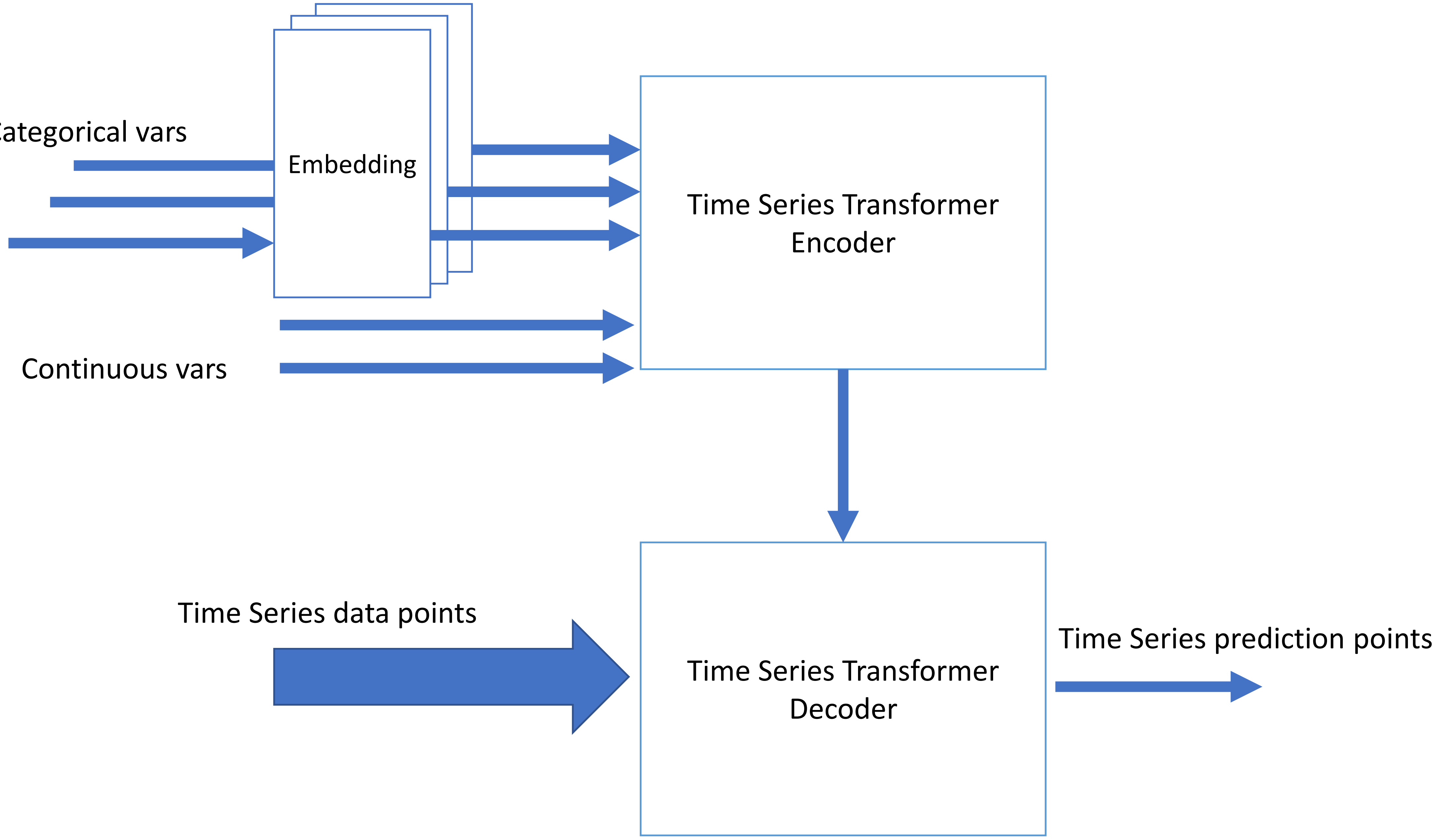


Fig: Architecture of Mixed Time Series Transformer

Critical Analysis

- Why not go with pure decoder architecture since it's just a next item prediction task?
  - Answer: Sticked with tried and tested model. Could potentially look into using a decoder only architecture.
- Bigger question on the inherent assumption i.e. "Delay of a segment is dependent upon past environment and past delays of the segment."
  - But realistically, delay also largely depends upon the past traffic conditions of ALL the other connected route segments. How do you encode that?

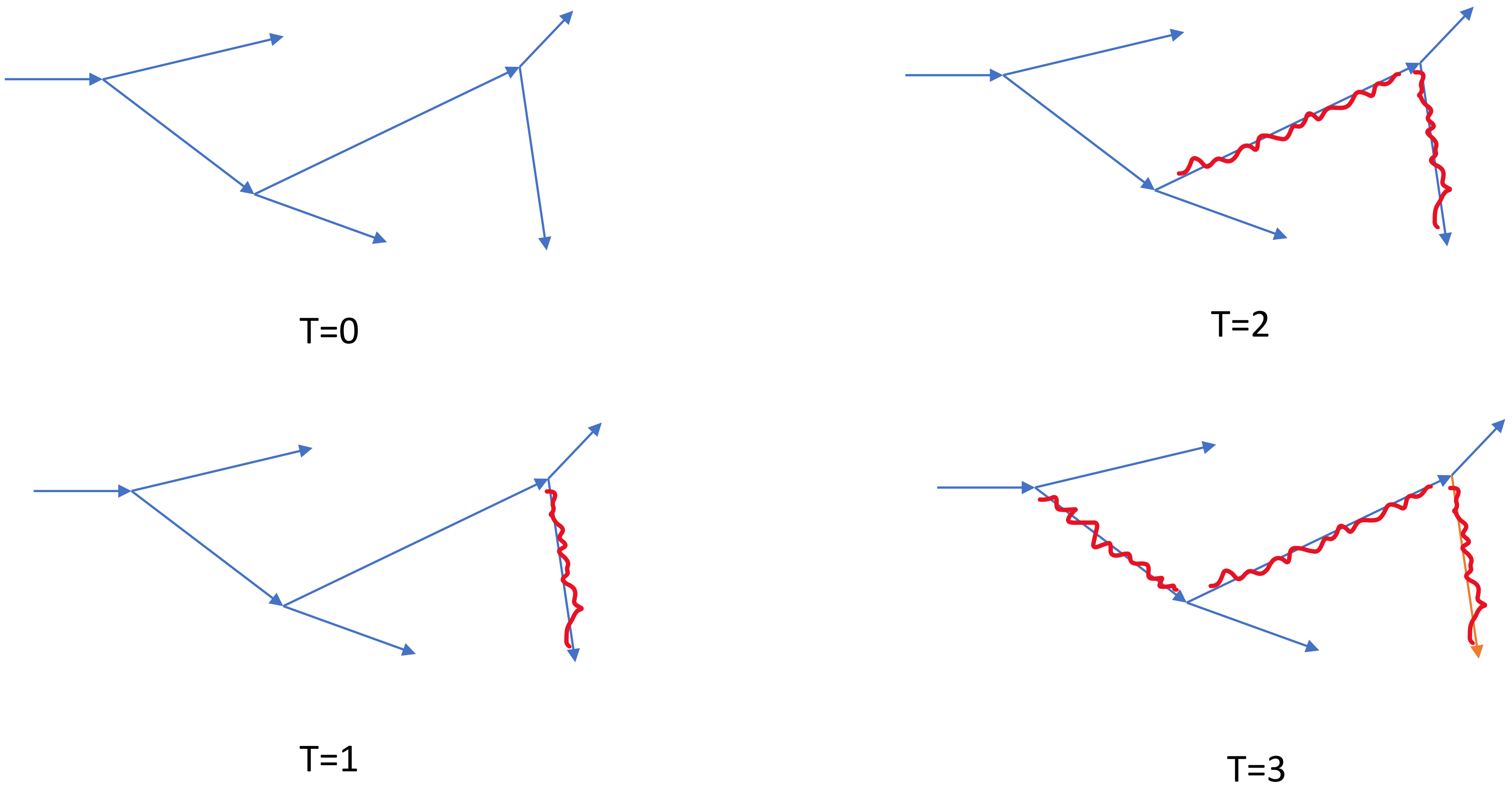


Fig: Delay propagation in Traffic Networks

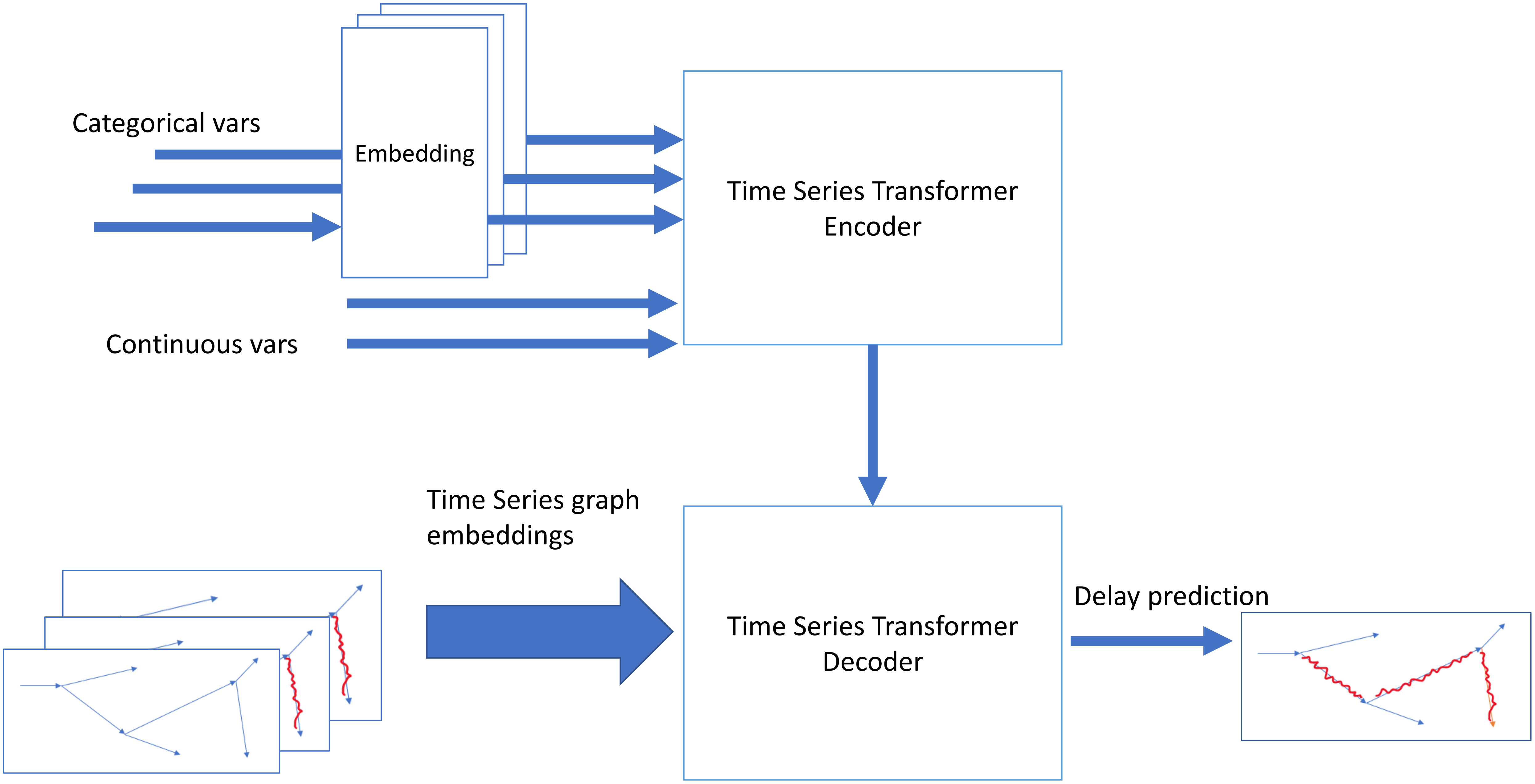


Fig: Architecture of Mixed Time Series Transformer

How do you incorporate the information of the global structure in Transformers?

- Generate Embedding using Graph Neural Networks?
- Predict $\mathbb{R}^n$ in decoder instead of a single $\mathbb{R}$?
  - Hoping that decoder self-attention will figure out a way to associate segment delays across time?