Algorithm for Time Series Forecasting

Input: $\mathbf{z}, \mathbf{x} \in \underline{\mathbb{R}^*}$, 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.

$\boldsymbol{W}_u \in \mathbb{R}^{1 \times d_e}$, the unembedding matrix.

$\boldsymbol{b}_u \in \mathbb{R}^1$, the unembedding bias.

/* Encoder portion */

$l_z \leftarrow length(\boldsymbol{z})$
for $t \in [l_z]$: $\boldsymbol{e}_t \leftarrow \boldsymbol{W}_e[:, z[t]] + \boldsymbol{W}_p[:, t]$
$\boldsymbol{Z} \leftarrow [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_{l_z}]$
for $l = 1,2,\dots,L_{enc}$ **do**
    $\boldsymbol{Z} \leftarrow \boldsymbol{Z} + MHAttention(\boldsymbol{Z} \,|\boldsymbol{W}_l^{enc}, Mask \equiv 1)$
    for $t \in [l_z]$: $\boldsymbol{Z}[:, t] \leftarrow layer\_norm(\boldsymbol{Z}[:, t] |\, \boldsymbol{\gamma}_l^1, \boldsymbol{\beta}_l^1)$
    $\boldsymbol{Z} \leftarrow \boldsymbol{Z} + \boldsymbol{W}_{mlp2}^l ReLU\left(\boldsymbol{W}_{mlp1}^l \boldsymbol{Z} + \boldsymbol{b}_{mlp1}^l \mathbf{1}^T\right) + \boldsymbol{b}_{mlp2}^l \mathbf{1}^{\boldsymbol{T}}$
    for $t \in [l_z]$: $\boldsymbol{Z}[:, t] \leftarrow layer\_norm(\boldsymbol{Z}[:, t] | \boldsymbol{\gamma}_l^2, \boldsymbol{\beta}_l^2)$
end

/* Decoder portion */

$l_x \leftarrow length(\boldsymbol{x})$
for $t \in [l_X]$: $\boldsymbol{e}_t \leftarrow \boldsymbol{W}_e[:, x[t]] + \boldsymbol{W}_p[:, t]$
$\boldsymbol{X} \leftarrow [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_{l_X}]$
for $l = 1,2,\dots,L_{dec}$ **do**
    $\boldsymbol{X} \leftarrow \boldsymbol{X} + MHAttention(\boldsymbol{X} | \boldsymbol{W}_l^{dec}, Mask[t, t'] \equiv [[t \leq t']])$
    for $t \in [l_X]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] |\, \boldsymbol{\gamma}_l^3, \boldsymbol{\beta}_l^3)$
    $\boldsymbol{X} \leftarrow \boldsymbol{X} + MHAttention(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{W}_l^{e/d}, Mask \equiv 1)$
    for $t \in [l_z]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] | \boldsymbol{\gamma}_l^4, \boldsymbol{\beta}_l^4)$
    $\boldsymbol{X} \leftarrow \boldsymbol{X} + \boldsymbol{W}_{mlp4}^l ReLU\left(\boldsymbol{W}_{mlp3}^l \boldsymbol{Z} + \boldsymbol{b}_{mlp3}^l \mathbf{1}^T\right) + \boldsymbol{b}_{mlp4}^l \mathbf{1}^{\boldsymbol{T}}$
    for $t \in [l_X]$: $\boldsymbol{X}[:, t] \leftarrow layer\_norm(\boldsymbol{X}[:, t] | \boldsymbol{\gamma}_l^5, \boldsymbol{\beta}_l^5)$
end
return $\boldsymbol{P} = \mathcal{F}(\boldsymbol{W_u X} + \boldsymbol{b\_u})$