# Appendix

## Formalization of Concept Bottleneck Framework

Any time series transformer obtains two types of input: (1) data values $\boldsymbol{X} \in \mathbb{R}^{I \times d}$, and (2) timestamps $\boldsymbol{T} \in \mathbb{R}^{I \times 4}$. The transformer consists of an encoder and a decoder, which are both constructed from one or multiple layers. Any encoder layer contains two sub-layers: a multi-head attention mechanism (Att) and a fully connected neural network (FF). Every sub-layer contains a residual connection around it. More specifically, the output $\boldsymbol{X}^\ell$ of any encoder layer $\ell$ is:

$$\begin{aligned} \boldsymbol{X}^\ell &= \text{Encoder}(\boldsymbol{X}^{\ell-1}) \\ &= \text{LayerNorm}(\text{FF}(\boldsymbol{S}^\ell) + \boldsymbol{S}^\ell), \\ \boldsymbol{S}^\ell &= \text{LayerNorm}(\text{Att}(\boldsymbol{X}^{\ell-1}) + \boldsymbol{X}^{\ell-1}), \end{aligned}$$

where

$$\begin{aligned} \text{FF}(\mathbf{x}) &= \max(0, \, \mathbf{x}\boldsymbol{W}_1 + \mathbf{b}_1) \, \boldsymbol{W}_2 + \mathbf{b}_2, \\ \text{Att}(\mathbf{x}) &= \boldsymbol{W}_0 \cdot \text{Concat}\left(\text{h}_1(\mathbf{x}), \, \ldots, \, \text{h}_h(\mathbf{x})\right). \end{aligned}$$

For future reference, we denote the output of the feed-forward module as follows: $\text{FF}(\boldsymbol{S}^\ell) = \boldsymbol{Z}^\ell \in \mathbb{R}^{d_1 \times d_2}$. We omit the definition of the decoder, because our bottleneck framework does not include it. Note that the exact implementation of each (sub-)layer depends on the type of transformer.

### *Concept Bottleneck Transformer Layer*

We assign one encoder layer to be the bottleneck and construct it such that it contains $c$ latent representations or *components*, i.e., $(\boldsymbol{H}_i)_{i=1}^c$. Depending on the bottleneck type $\tau$, these latent representations are either taken from the attention mechanism or the feed-forward module. More specifically:

$$\boldsymbol{H}_i = \begin{cases} \text{h}_i(\mathbf{x}) & \text{if bottleneck type } \tau = \text{Att}, \\ \boldsymbol{Z}_i & \text{if bottleneck type } \tau = \text{FF}. \end{cases}$$

Since the attention block is multi-headed, different heads naturally form the components of the attention bottleneck. For the feed-forward bottleneck, we define the components to be slices (in $d_1$) from its output $\boldsymbol{Z}$, such that stacking the components results in the original output.

Note that the residual connection around the corresponding bottleneck component is removed, and that each component $\boldsymbol{H}_i$ should represent a predefined interpretable concept.

### *Intervention*

In the intervention experiment, we shift the time stamps $\boldsymbol{T}$ to obtain $\widetilde{\boldsymbol{T}}$. The key aspect of the experiment is to run the transformer on the shifted time stamps $\widetilde{\boldsymbol{T}}$, and replace the input representations $\widetilde{\boldsymbol{X}}^{b-1}$ of the bottleneck layer $b$ with $\boldsymbol{X}^{b-1}$ (based on $\boldsymbol{T}$), but only in the component that represents the time concept.

More specifically, if type $\tau = \text{Att}$, we intervene on the attention block in the bottleneck as follows:

$$\text{Att}(\mathbf{x}, \widetilde{\mathbf{x}}) = \boldsymbol{W}_0 \cdot \text{Concat}\left(\mathrm{h}_1(\widetilde{\mathbf{x}}),\, \mathrm{h}_2(\mathbf{x}),\, \mathrm{h}_3(\widetilde{\mathbf{x}})\right),$$

and, if type $\tau = \text{FF}$, as follows:

$$\text{FF}(\mathbf{x}, \widetilde{\mathbf{x}}) = \text{Stack}(\widetilde{\boldsymbol{Z}}_1, \boldsymbol{Z}_2, \widetilde{\boldsymbol{Z}}_3).$$

In both functions we make use of the fact that the time concept is represented in the second component, and there are three components in total. This intervention can be done in the bottleneck only, because, by construction, the location of its concept representations is known.

## Detailed results

**Table 3.** Performance of different models in Mean Squared Error (MSE) and Mean Absolute Error (MAE). The bottlenecks **do** contain a free component ($c = 3$), and use AR as surrogate model. The model with no bottleneck is an original Autoformer of similar size. For all datasets, the shortest prediction lengths from Wu et al. (2021) are used, see Table 1. The standard deviation is determined using five different seeds.

| Free component | Att bottleneck | | FF bottleneck | | No bottleneck | | AR | | Wu et al. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| Electricity | 0.231 ± 0.009 | 0.338 ± 0.005 | 0.207 ± 0.005 | 0.320 ± 0.005 | 0.280 ± 0.165 | 0.368 ± 0.111 | 0.497 | 0.522 | **0.201** ± 0.003 | **0.317** ± 0.004 |
| Traffic | 0.642 ± 0.022 | 0.393 ± 0.013 | **0.393** ± 0.013 | **0.377** ± 0.006 | 0.619 ± 0.015 | 0.387 ± 0.005 | 0.420 | 0.494 | 0.613 ± 0.028 | 0.388 ± 0.012 |
| Weather | 0.290 ± 0.027 | 0.354 ± 0.020 | 0.271 ± 0.016 | 0.341 ± 0.011 | 0.269 ± 0.000 | 0.344 ± 0.000 | **0.006** | **0.062** | 0.266 ± 0.007 | 0.336 ± 0.006 |
| Illness | 3.586 ± 0.241 | 1.313 ± 0.040 | 3.661 ± 0.237 | 1.322 ± 0.050 | 3.405 ± 0.208 | 1.295 ± 0.044 | **1.027** | **0.820** | 3.483 ± 0.107 | 1.287 ± 0.018 |
| Exchange rate | 0.195 ± 0.029 | 0.323 ± 0.025 | 0.155 ± 0.010 | 0.290 ± 0.013 | 0.152 ± 0.003 | 0.283 ± 0.003 | **0.082** | **0.230** | 0.197 ± 0.019 | 0.323 ± 0.012 |
| ETT | 0.177 ± 0.003 | 0.282 ± 0.004 | 0.174 ± 0.006 | 0.280 ± 0.005 | 0.155 ± 0.004 | 0.265 ± 0.002 | **0.034** | **0.117** | 0.255 ± 0.020 | 0.339 ± 0.020 |

**Table 4.** Performance on different datasets, where the bottlenecks **do not** contain a free component **(**$c = 2$**)**. AR is used as surrogate model in the bottlenecks. The model with no bottleneck is an original Autoformer of similar size. For all datasets, the shortest prediction lengths from Wu et al. (2021) are used, see Table 1. The standard deviation is determined using five different seeds.

| No free component | Att bottleneck | | FF bottleneck | | No bottleneck | | AR | | Wu et al. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| Electricity | 0.224 ± 0.006 | 0.332 ± 0.003 | 0.206 ± 0.009 | 0.321 ± 0.009 | 0.202 ± 0.006 | 0.318 ± 0.007 | 0.497 | 0.522 | **0.201** ± 0.003 | **0.317** ± 0.004 |
| Traffic | 0.629 ± 0.023 | 0.394 ± 0.015 | 0.627 ± 0.031 | 0.392 ± 0.025 | 0.613 ± 0.018 | **0.378** ± 0.007 | **0.420** | 0.494 | 0.613 ± 0.028 | 0.388 ± 0.012 |
| Weather | 0.281 ± 0.025 | 0.348 ± 0.018 | 0.260 ± 0.015 | 0.333 ± 0.013 | 0.257 ± 0.004 | 0.332 ± 0.005 | **0.006** | **0.062** | 0.266 ± 0.007 | 0.336 ± 0.006 |
| Illness | 3.966 ± 0.296 | 1.401 ± 0.073 | 3.721 ± 0.268 | 1.351 ± 0.053 | 3.585 ± 0.331 | 1.333 ± 0.070 | **1.027** | **0.820** | 3.483 ± 0.107 | 1.287 ± 0.018 |
| Exchange rate | 0.208 ± 0.026 | 0.333 ± 0.022 | 0.158 ± 0.009 | 0.293 ± 0.009 | 0.152 ± 0.006 | 0.284 ± 0.007 | **0.082** | **0.230** | 0.197 ± 0.019 | 0.323 ± 0.012 |
| ETT | 0.178 ± 0.011 | 0.283 ± 0.007 | 0.174 ± 0.01 | 0.283 ± 0.009 | 0.165 ± 0.004 | 0.274 ± 0.004 | **0.034** | **0.117** | 0.255 ± 0.020 | 0.339 ± 0.020 |