Adjustable Context-aware Transformers

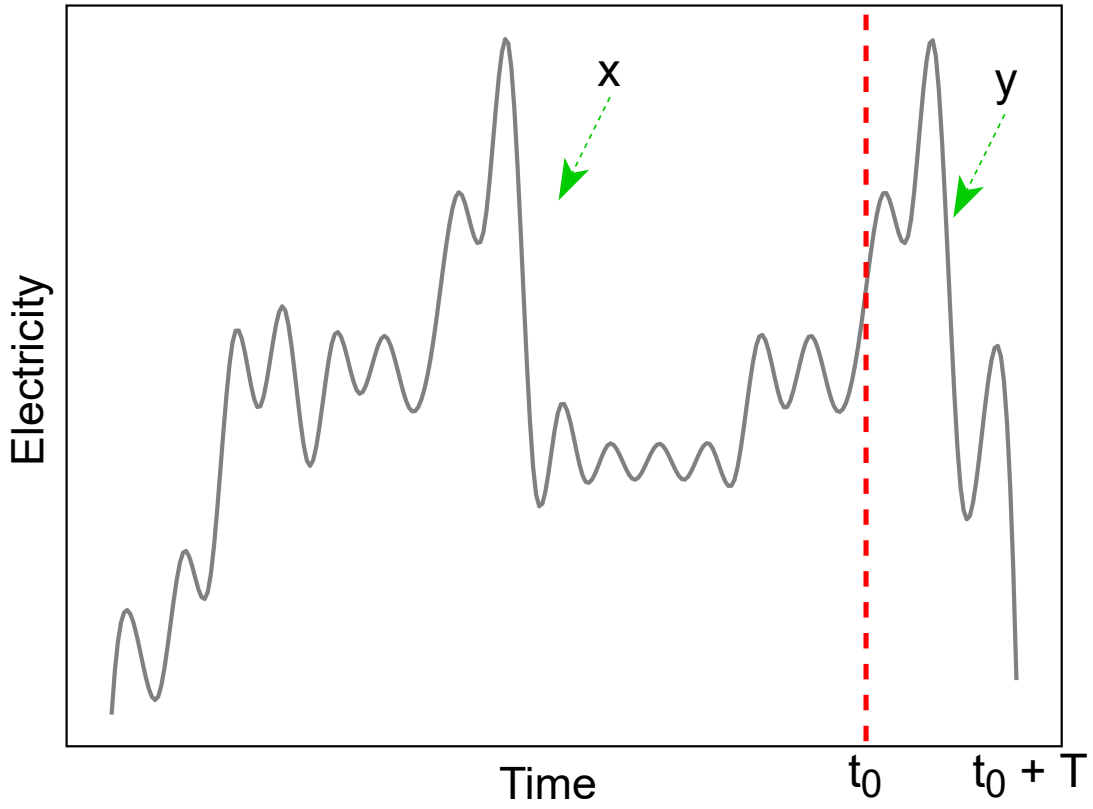
Presenter: Sepideh Koohfar

AALTD@ECML 2022

Problem Definition

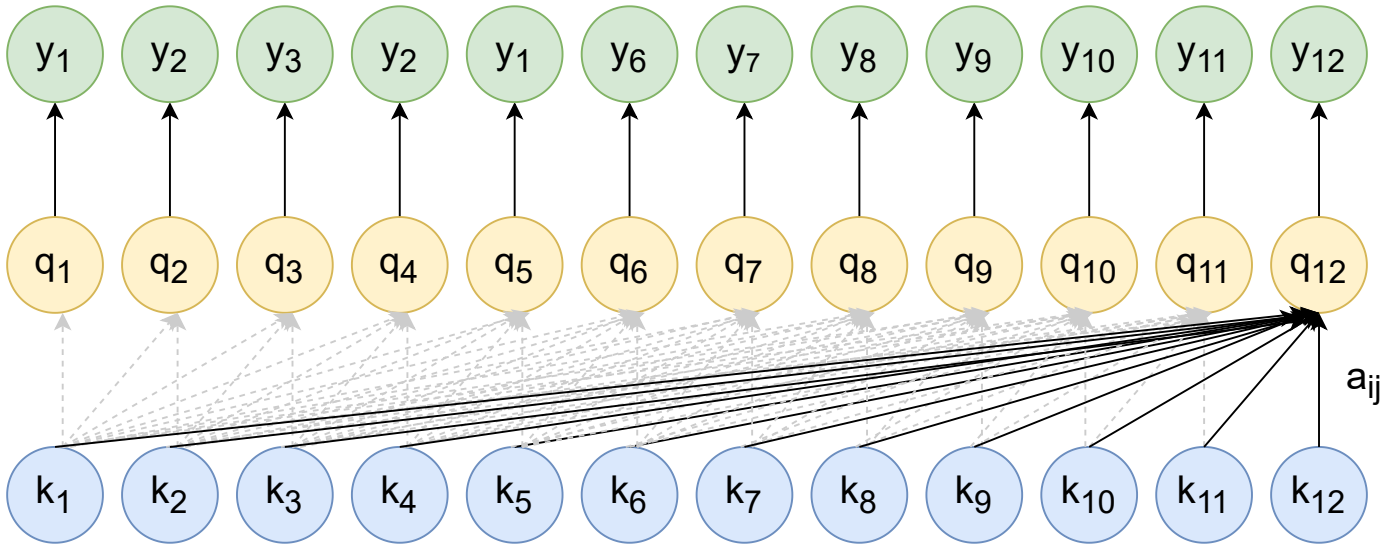
- Time series forecasting is a vital problem: climate, weather, energy
- Multi-horizon forecasting is a critical demand: early sever weather events forecasting

- Given the input data prior to time step $t_{0,}$ the task is to predict the variables of interest for multiple steps into the future from t_0 to t_0+T.



Background: Transformers

- Transformers can incorporate any observations of the series; it renders them more suitable for capturing similarities in the longer past.

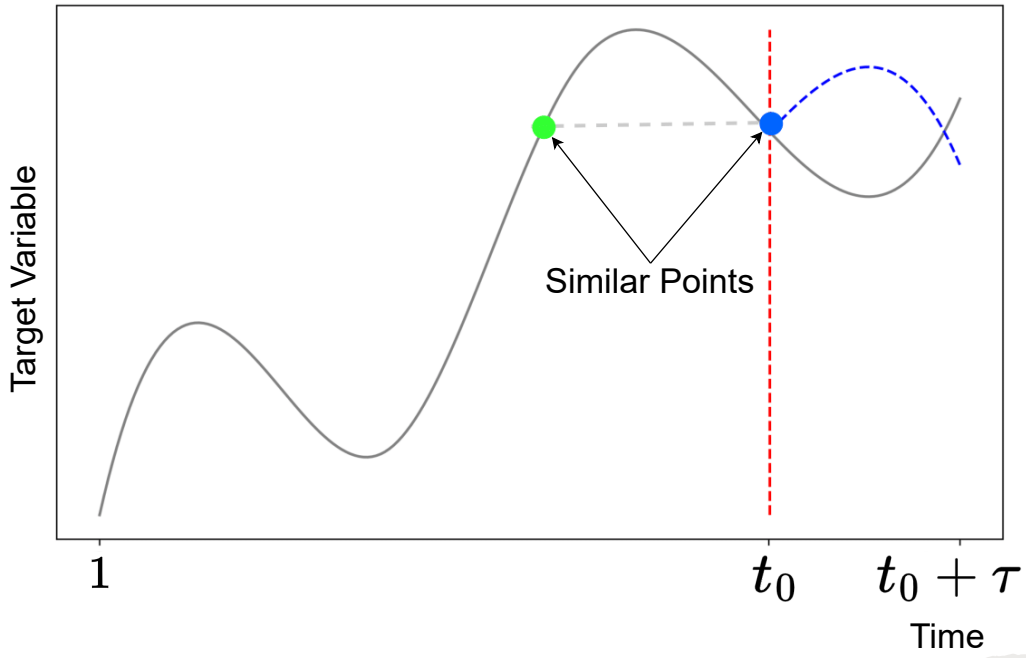


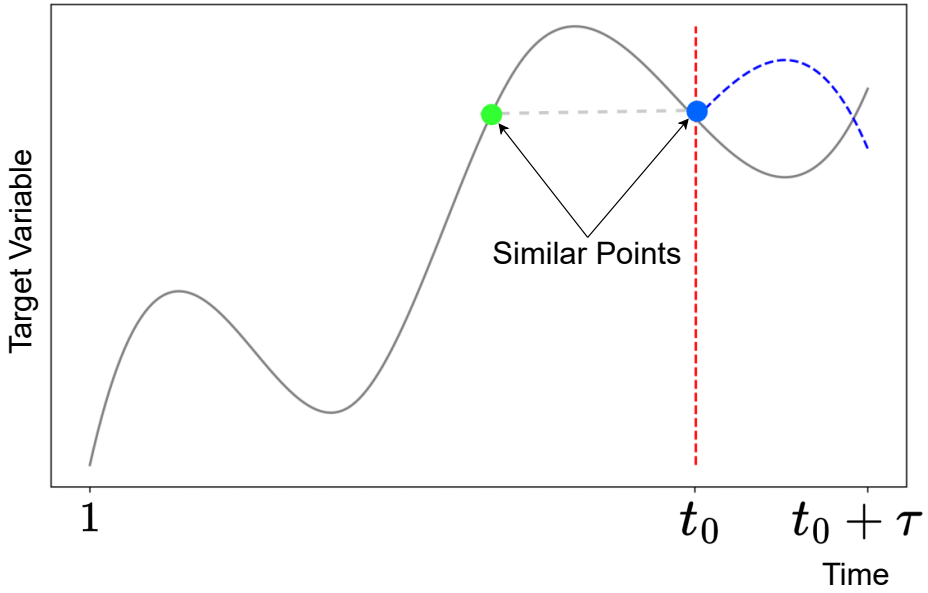
$q_i = \text{proj}(x_i)$, $k_j = \text{proj}(x_j)$

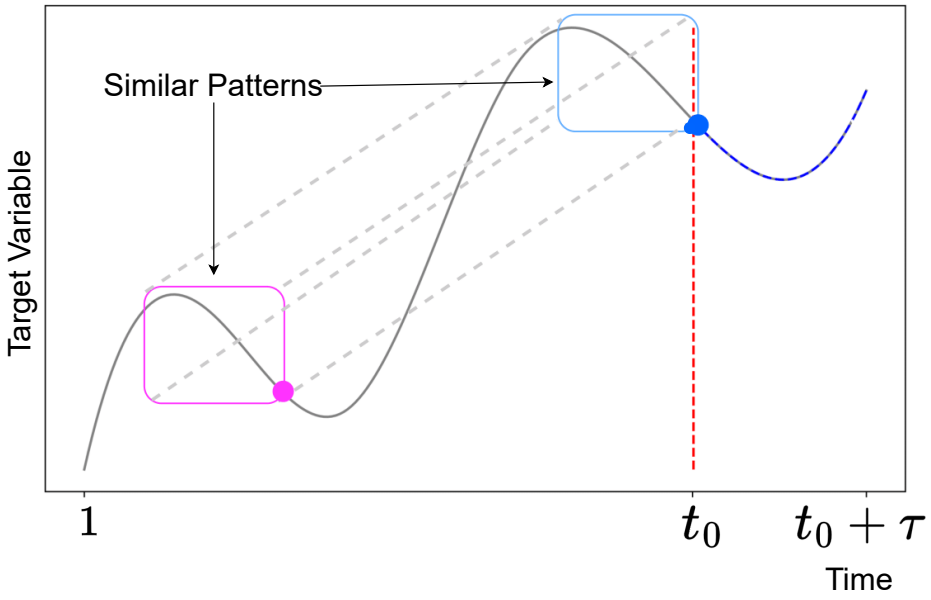
$$y_i = \sum_{j \le i} a_{ij} x_j$$

Background:
Issues arising from point-wise attention

- Point-wise attention only considers the information at time step i and j

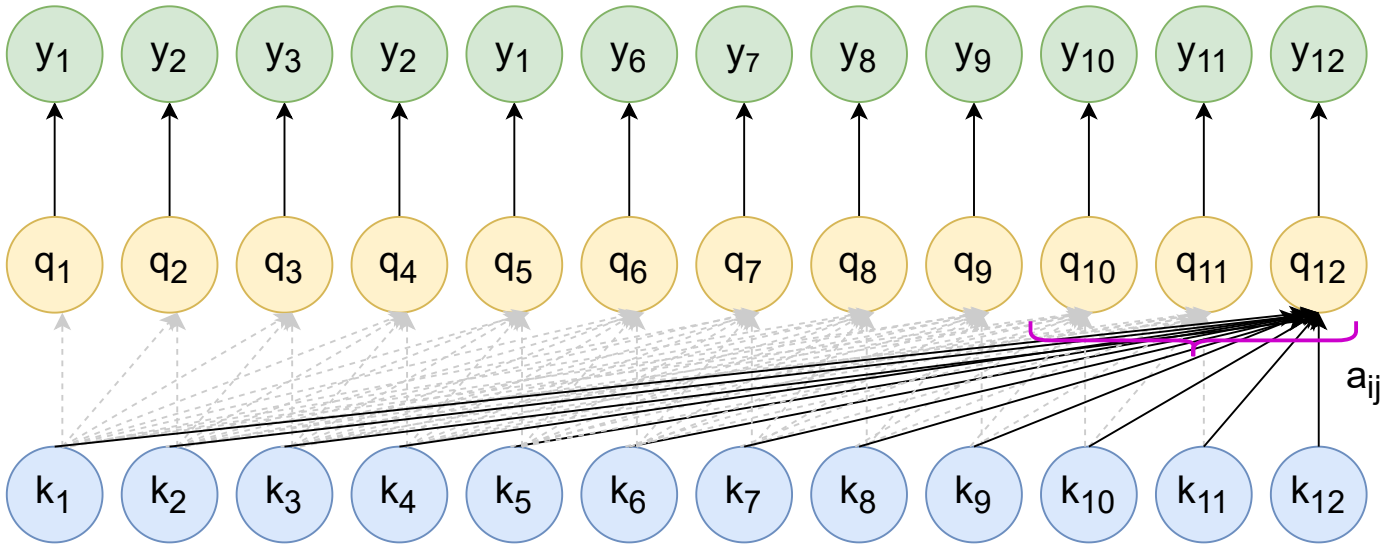



Target Variable
Similar Points
1
t_0
$t_0 + \tau$
Time


Similar Patterns
Target Variable
1
t_0
$t_0 + \tau$
Time

Temporal Attention

- Temporal attention is achieved by deriving query and key vectors from a temporal context with length w preceding time step i



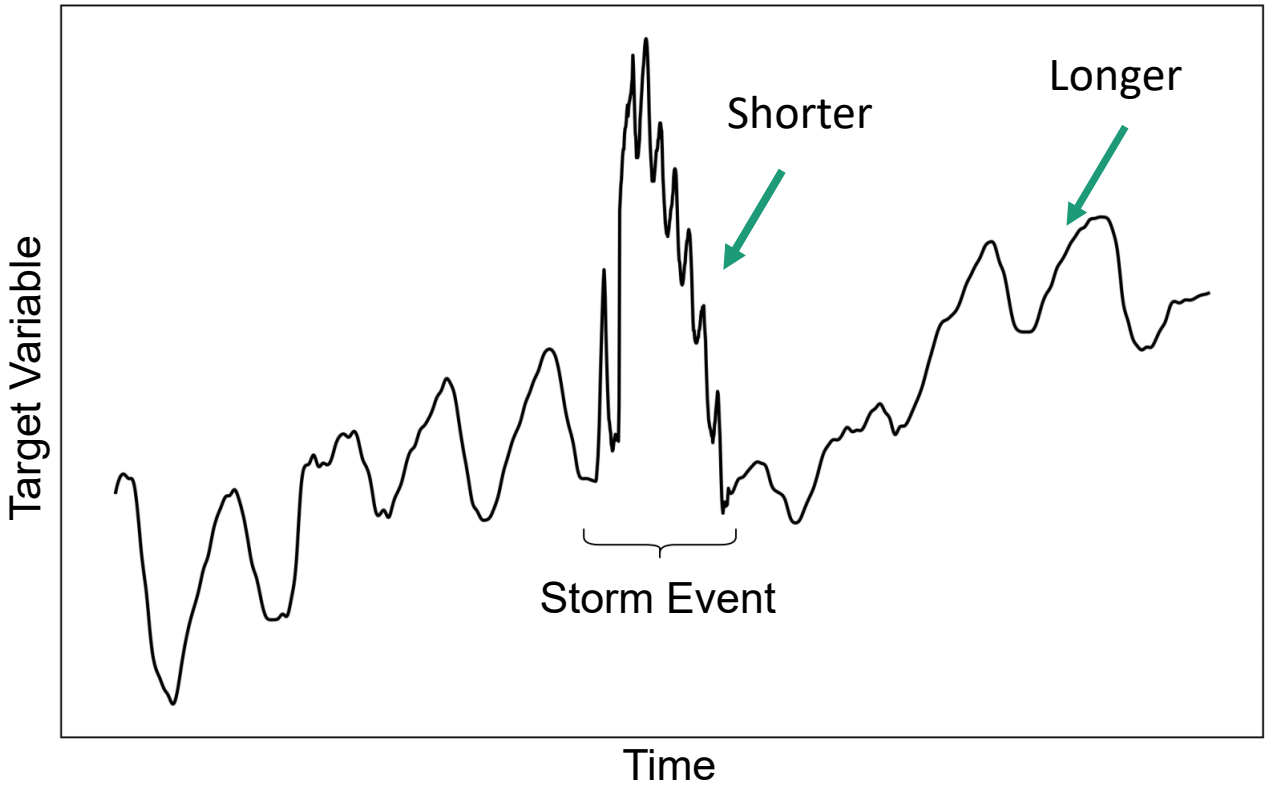
$$q_i = proj(x_{i-3:i}) \quad k_j = proj(x_{j-3:j})$$

One Limitation of Temporal Attention

- Context length is of a fixed size
- A noise introduced by an excessive large context length can be misleading
- We propose an alternative for this problem to choose an optimal context length for each query and key

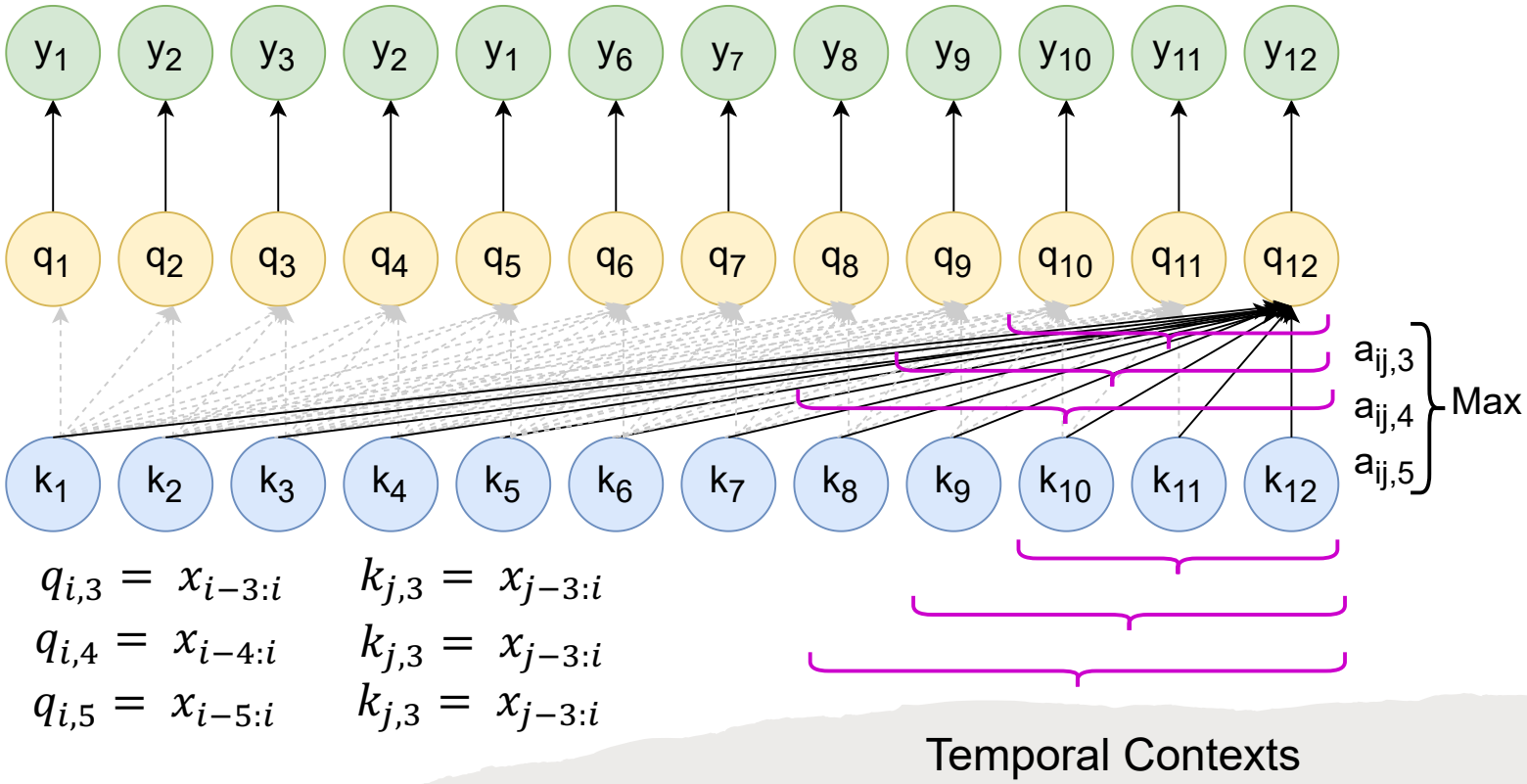
Adjustable Context-aware Attention

- We hypothesize that a successful needs to switch between different lengths dynamically depending on the situation.



Adjustable Context-aware Transformer

- We consider multiple context when computing the attention score and make the selection of the ideal context length part of the prediction problem using the following model:

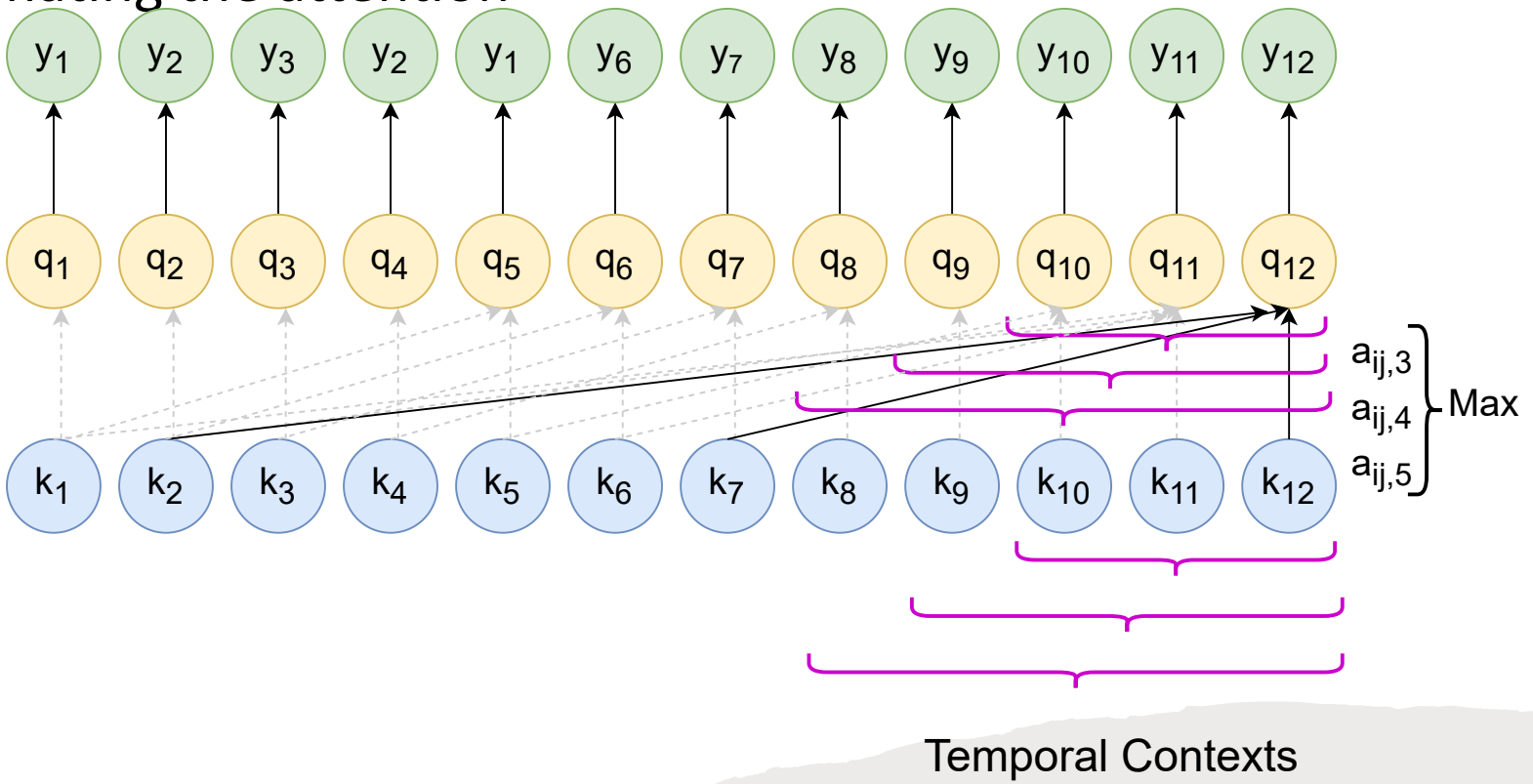


Adjustable Context-aware Transformer

- The downside is the demand of resources
- Our model needs to explore all possible context-aware key and key pair
- For an attention model with Q queries, K keys, and W context lengths, computing the attention weights requires $\mathcal{O}(W.Q.K)$ space and time.

(Efficient) Adjustable Context-aware Transformer

- We propose to subsample the time points for estimating the attention scores



Experiments: Datasets

- Electricity: The univariate Electricity Load Diagrams dataset, containing the electricity consumption of 370 customers, hourly level.
- Traffic: A Univariate dataset containing occupancy rate of 440 SF Bay Area, hourly level.
- Watershed: This multivariate dataset contains hydrological streamflow responses of ten watershed sites, 15 minutes level.

Experiments: Baselines

- ARIMA: Auto-regressive integrated moving average.
- LSTM: Long short-term memory networks
- Transformer: A single layer transformer equivalent to our approach with the basic multi-head attention.
- Trans-multi: A three encoder layer and one decoder layer transformer with multi-head basic attention.
- CNN-trans: A single layer transformer with convolutional multi-head attention.

Main Results

	Horizon	Metric	ARIMA	LSTM	Transformer	Trans-multi	CNN-trans	ACAT (Ours)
Traffic	24	NRMSE	0.805 ± 0.0000	0.502 ± 0.0050	0.475 ± 0.0003	0.586 ± 0.0214	0.474 ± 0.0001	**0.375** ± 0.0002 (+**21**%)
		NMAE	0.559 ± 0.0000	0.279 ± 0.0052	0.245 ± 0.0008	0.352 ± 0.0212	0.243 ± 0.0006	**0.157** ± 0.0003 (+**35**%)
	48	NRMSE	0.794 ± 0.0000	0.485 ± 0.0010	0.458 ± 0.0001	0.680 ± 0.0213	0.455 ± 0.0002	**0.354** ± 0.0002 (+**22**%)
		NMAE	0.559 ± 0.0000	0.264 ± 0.0009	0.237 ± 0.0003	0.452 ± 0.0204	0.233 ± 0.0003	**0.156** ± 0.0005 (+**33**%)
Electricity	24	NRMSE	3.984 ± 0.0000	1.286 ± 0.0076	1.292 ± 0.0171	1.484 ± 0.0193	1.265 ± 0.0185	**0.638** ± 0.0050 (+**50**%)
		NMAE	0.409 ± 0.0000	0.128 ± 0.0010	0.148 ± 0.0013	0.163 ± 0.0010	0.143 ± 0.0012	**0.080** ± 0.0002 (+**38**%)
	48	NRMSE	4.055 ± 0.0000	1.401 ± 0.0154	1.467 ± 0.0051	1.562 ± 0.0326	1.263 ± 0.0087	**0.843** ± 0.0078 (+**33**%)
		NMAE	0.412 ± 0.0000	0.141 ± 0.0016	0.159 ± 0.0002	0.166 ± 0.0013	0.143 ± 0.0005	**0.091** ± 0.0002 (+**35**%)
Watershed	24	NRMSE	-	0.353 ± 0.0112	0.332 ± 0.0028	0.345 ± 0.0031	0.340 ± 0.0005	**0.283** ± 0.0012 (+**15**%)
		NMAE		0.202 ± 0.0076	0.193 ± 0.0023	0.208 ± 0.0014	0.199 ± 0.0002	**0.156** ± 0.0015 (+**19**%)
	48	NRMSE	-	0.418 ± 0.0103	0.356 ± 0.0027	0.346 ± 0.0015	0.352 ± 0.0021	**0.309** ± 0.0015 (+**11**%)
		NMAE		0.252 ± 0.0076	0.206 ± 0.0023	0.203 ± 0.0014	0.201 ± 0.0017	**0.164** ± 0.0010 (+**18**%)

Table 1: Results summary in NRMSE and NMAE of all methods on three datasets. Best results are highlighted in boldface. The loss of our model is significantly lower compared to other baselines in both evaluation metrics.

The Importance of Our Model

	Horizon	Metric	ACAT (Ours)	$l = 1$ TA	$l = 3$ TA	$l = 6$ TA	$l = 9$ TA
Traffic	24	NRMSE	**0.375** ± 0.0002	0.487 ± 0.0003	0.482 ± 0.0008	0.475 ± 0.0003	0.476 ± 0.0006
		NMAE	**0.157** ± 0.0003	0.260 ± 0.0009	0.251 ± 0.0012	0.247 ± 0.0002	0.245 ± 0.0011
	48	NRMSE	**0.354** ± 0.0002	0.462 ± 0.0006	0.459 ± 0.0005	0.463 ± 0.0005	0.460 ± 0.0006
		NMAE	**0.156** ± 0.0005	0.242 ± 0.0005	0.236 ± 0.0005	0.243 ± 0.0006	0.237 ± 0.0005
Electricity	24	NRMSE	**0.638** ± 0.0050	1.295 ± 0.0043	1.178 ± 0.0102	1.382 ± 0.0145	1.278 ± 0.0071
		NMAE	**0.080** ± 0.0002	0.147 ± 0.0001	0.141 ± 0.0006	0.148 ± 0.0007	0.142 ± 0.0004
	48	NRMSE	**0.843** ± 0.0078	1.325 ± 0.0047	1.394 ± 0.0098	1.397 ± 0.0163	1.263 ± 0.0104
		NMAE	**0.091** ± 0.0002	0.147 ± 0.0003	0.150 ± 0.0006	0.152 ± 0.0009	0.146 ± 0.0008
Watershed	24	NRMSE	**0.283** ± 0.0012	0.360 ± 0.0033	0.370 ± 0.0065	0.337 ± 0.0027	0.367 ± 0.0027
		NMAE	**0.156** ± 0.0015	0.215 ± 0.0038	0.222 ± 0.0054	0.194 ± 0.0025	0.220 ± 0.0019
	48	NRMSE	**0.309** ± 0.0015	0.354 ± 0.0026	0.340 ± 0.0025	0.384 ± 0.0024	0.376 ± 0.0015
		NMAE	**0.164** ± 0.0010	0.204 ± 0.0009	0.192 ± 0.0018	0.229 ± 0.0018	0.225 ± 0.0010

Table 2: Comparison of fixed and adjustable context length approaches. TA represents a temporal transformer with a fixed context size. The context length l is set as a hyperparamteter. Best results are highlighted in boldface. We can observe that not adjusting the context length drastically affect the performance.

Conclusion

- We propose ACAT, the adjustable context-aware transformer.
- ACAT automatically selects the ideal context size to obtain the best forecasting results.
- ACAT model obtains performance improvements over state-of-the-art temporal attention approaches.
- This indicates that incorporating the ideal context length in the query-key similarity of the attention mechanism can improve the forecasting quality.

References

1. Ashish Vaswani et al. “Attention is All you Need”. In: Advances in Neural Information Processing Systems. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017
2. Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 471, 5243–5253.