AttentionPredictor

Why -> Because i was fascinated by the fact that there exist the solution to KV cache storage problem bottleneck (KV cache compression)

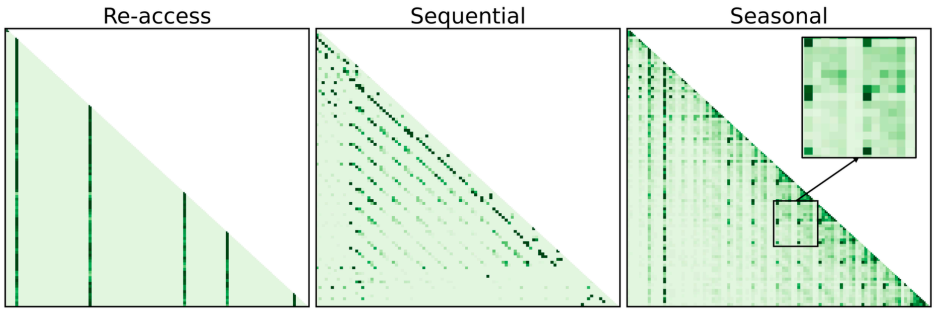


Figure 2: Visualization of three temporal attention patterns. **Re-access** shows repeated attention to specific tokens. **Sequential** shows attention progresses toward the next tokens. **Seasonal** exhibits periodic recurrence as alternating bands of high and uniform attention scores.

Observation ^^^

Mathematical proof that attention of (t+1)th time is related to 't'th time ones / why attention follows temporal patterns ?

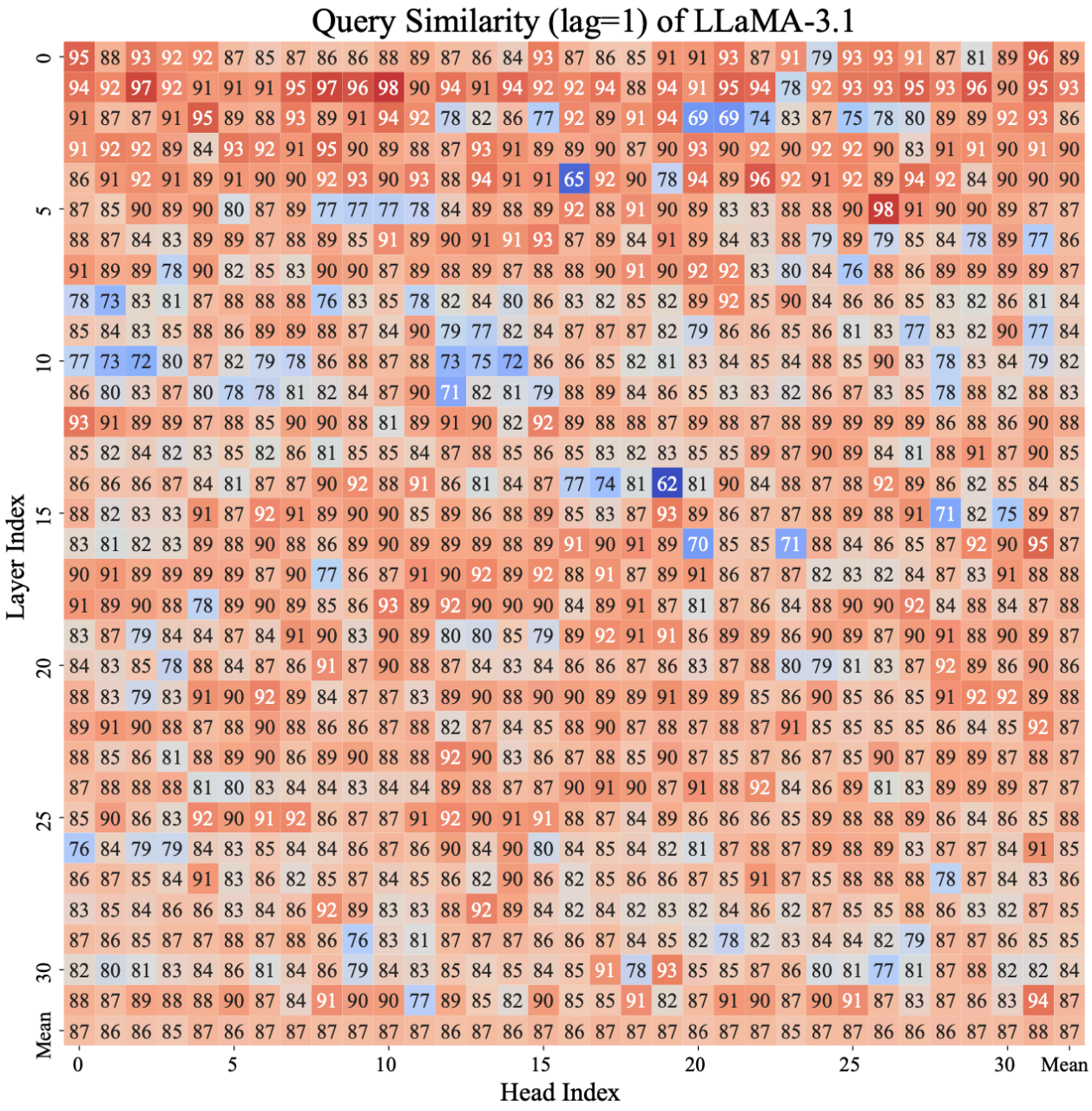
$$
\begin{aligned}
A_{i+1} &= \frac{1}{\sqrt{D}} \mathbf{Q_{i+1}} \mathbf{K_{i+1}}^{\top} \\
&= \frac{1}{\sqrt{D}} q_{i+1} k_{1:i+1}^{\top} \\
&= \frac{1}{\sqrt{D}} (q_i + \Delta q) k_{1:i+1}^{\top} \\
&= \frac{1}{\sqrt{D}} (q_i k_{1:i+1}^{\top} + \Delta q k_{1:i+1}^{\top})
\end{aligned}
$$

Focusing on the values of A_{i+1} in the first i positions,

$$A_{i+1}[1:i] = \frac{1}{\sqrt{D}} q_i k_{1:i}^\top + \frac{1}{\sqrt{D}} \Delta q k_{1:i}^\top$$
$$= A_i + \Delta A$$

why q_i+1 = q_i + delta (q) means why the next query is almost similar to previous token's similarity ?

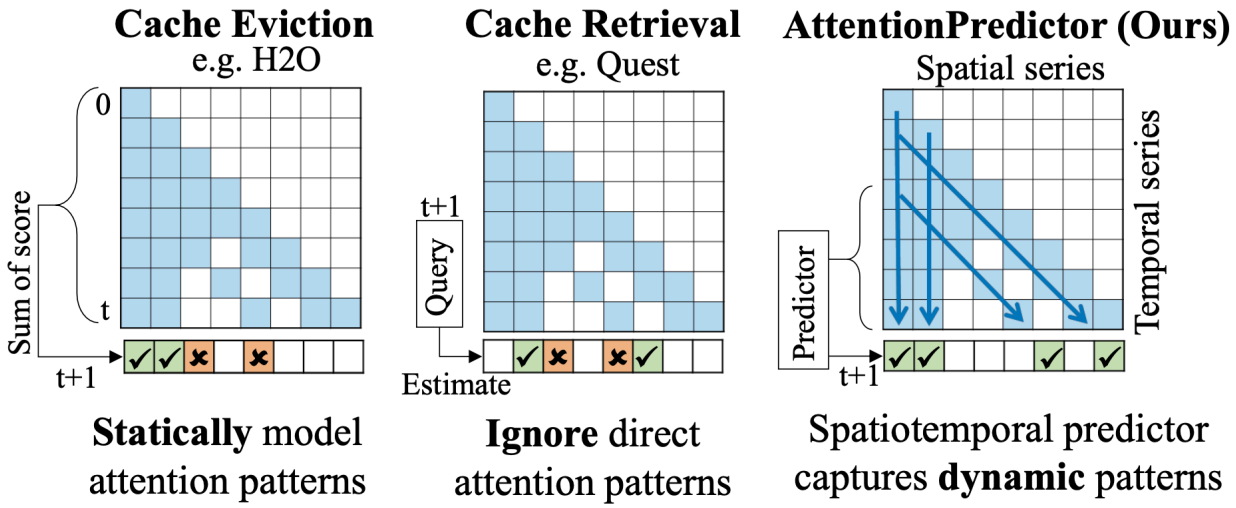
`the query sequence exhibits a cosine autocorrelation of 87% at a one-step lag,which highlights its strong similarity.`



based on this observation what happened is we can conclude that the attention scores are kind of a time series data so they can be predicted

So attentionPredictor took benefit of this thing and created a time series model that will predict the next token attention score prediction and based on that prediction what we can do is select only the critical KV tokens to save the compute and memory

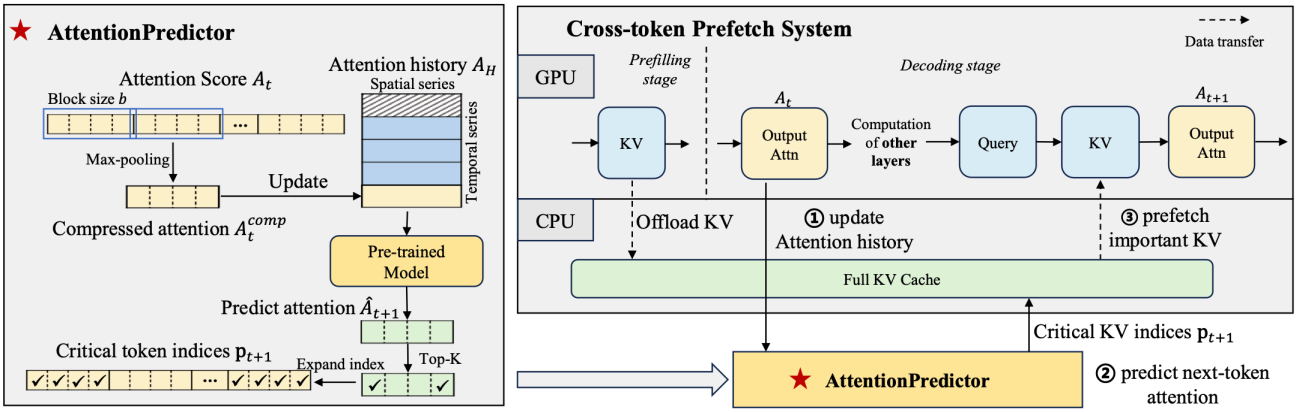
Was it tried before ?
yeah



Previously, two major ideas were explored: one focused on selecting the most critical token based on the highest attention score, while the other was a retrieval-based approach. However, the retrieval-based method failed for longer contexts and also the retrieval based method can not be efficiently updated as it needed the current query token

So thats why as attention predictor is taking care of temporal series it was superior than the previous methods

How the model was implemented ?



it used CNN as it was lightweight than the LSTM's , also able to capture the spatial relationship

One more thing that was implmented that bit on the engineering side was the Cross-token Prefetch System which before hand predicts the next attention score and based on that the critical tokens which were than loaded to GPU from CPU instead of all the KV tokens which was the main key idea that will help to icrease the context length and save memory and compute.

there are more details about the pooling,dense KV cache for calibration and other stuff but i covered the main ideas

| AttentionPredictor, the first learning-based critical token identification approach for KV cache compression.
| AttentionPredictor achieves comparable accuracy of LLM with 16×KV cache compression.