Towards Out-of-Distribution Sequential Event Prediction: A Causal Treatment

A Comparative Analysis and Proposed Enhancement

1/31/25

Introduction

The paper addresses the challenge of sequential event prediction where models are trained on historical sequences and must generalize to future and unseen events:

- Objective: Addressing challenges in sequential event prediction under temporal distribution shifts
- Key Problem: Traditional models fail to generalize when future data distributions differ from training data.
- Solution Space: CaseQ framework and a proposed alternative solution.

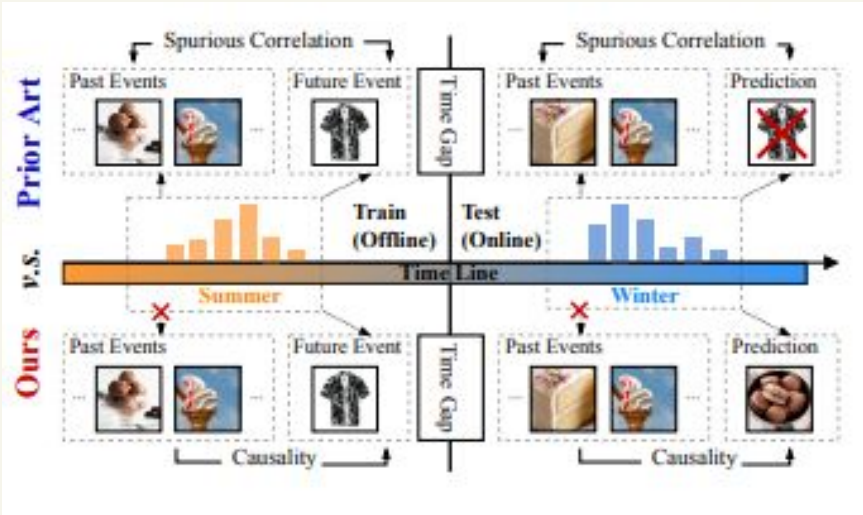
Understanding Sequential Event Prediction

- Sequential Event Prediction: Predicting the next event given a historical sequence of past events.

- Applications:
  - Recommendation systems
  - User behavior modeling
  - Clinical treatment predictions
  - Industrial maintenance

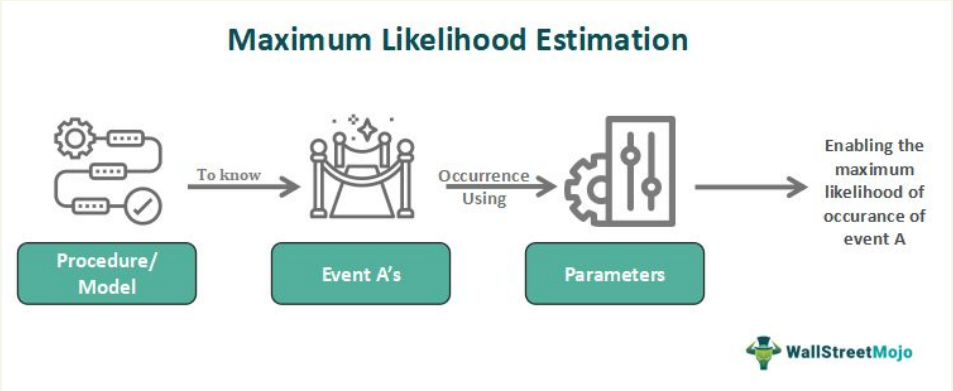
- Challenges: Temporal shifts cause models trained on past data to fail in future scenarios.

- Figure 1: A toy example in recommendation. Prior art would spuriously correlate non-causal items (‘ice cream’ and ‘beach T-shirt’) and produce undesired results under a new environment in the future.



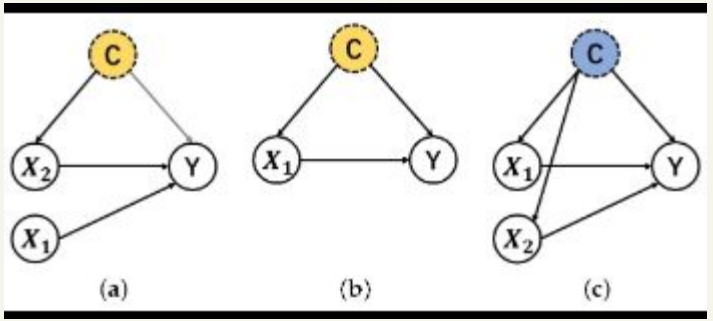
Limitations of Traditional Approaches

- Maximum Likelihood Estimation based models (MLE): Optimizes the likelihood of observed sequences
    - Issue: They learn spurious correlations and fail under temporal shifts
-
    - Example: A recommender system learns that users who buy ice cream also buy beach T-Shirt (seasonal bias).
    - Root Cause: Latent context confounders (external factors that influence both past and future events).



Variational Context Adjustment

- Variational Context Adjustment:
  - They propose a backdoor adjustment method using variational inference to account for the confounding effect of external contexts
  - This method helps models learn true causal relationships rather than relying on spurious correlations.



Introducing CaseQ

CaseQ: A hierarchical sequence model that learns context-aware representation and uses a branching mechanism to adapt to different contexts dynamically.

1.) Event Embedding Layer
   a.) Each event x_m is mapped to a dense vector using a global embedding matrix H_x.

2.) Context-Specific Encoder Φ
   a.) Instead of a single RNN or Transformer, they introduce context-aware encoders Φ_k.
      i.) $h_t = \Sigma_k c_t[k] * \Phi_k(S)$
   b.) The encoder dynamically switches between different context-specific sub-models.

Introducing CaseQ (Continued)

1.) Branching Unit Ψ for Context Estimation
   a.) Learns probabilities of different contexts using **softmax**.
      i.) $Q_t = \text{Softmax}([\langle a_k, \tanh(W_k h_t)\rangle]_k, \tau)$
   b.) Uses Gumbel-Softmax to make context selection differentiable

2.) Hierarchical Context Representation
   a.) Instead of a single-layer model, they use a hierarchical architecture where contexts are learned across multiple layers:

   b.) This allows for generalization to unseen contexts.

Experimental Results

	Gap Size	GRU4Rec [20]		CaseQ (GRU)		SASRec [24]		CaseQ (SA)		SSE-PT [49]		CaseQ (PT)	
		NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR
Movielens	0	0.436	0.730	0.448 (+1.1%)[1]	0.736 (+0.8%)	0.453	0.742	0.462 (+2.0%)	0.752 (+1.3%)	0.465	0.753	0.476 (+2.3%)	0.763 (+1.3%)
	30	0.371	0.603	0.406 (+8.7%)	0.660 (+8.7%)	0.384	0.627	0.429 (+10.5%)	0.675 (+7.1%)	0.397	0.642	0.443 (+10.3%)	0.680 (+5.5%)
	Drop (%)	↓ 15.0	↓ 17.4	↓ 9.3 (−38.5%)[2]	↓ 10.3 (−40.7%)	↓ 15.2	↓ 15.6	↓ 7.1 (−53.1%)	↓ 9.3 (−40.6%)	↓ 14.6	↓ 14.7	↓ 7.0 (−51.8%)	↓ 10.9 (−25.9%)
Yelp	0	0.428	0.714	0.440 (+2.8%)	0.731 (+2.3%)	0.436	0.734	0.452 (+3.7%)	0.744 (+1.4%)	0.451	0.745	0.462 (+2.3%)	0.751 (+0.9%)
	20	0.395	0.663	0.426 (+7.2%)	0.702 (+5.5%)	0.402	0.682	0.433 (+7.0%)	0.719 (+5.2%)	0.405	0.685	0.440 (+8.0%)	0.717 (+4.5%)
	Drop (%)	↓ 7.6	↓ 7.2	↓ 3.2 (−57.9%)	↓ 4.1 (−43.1%)	↓ 7.6	↓ 7.1	↓ 4.3 (−43.3%)	↓ 3.4 (−52.5%)	↓ 10.3	↓ 8.0	↓ 4.7 (−54.5%)	↓ 4.5 (−43.7%)

[1] The percentage measures the improvement of CaseQ over the counterpart baseline with respect to one metric (e.g., NDCG).
[2] The percentage measures the degree of CaseQ helping to alleviate the performance drop compared to the counterpart.

Table 1: Experiment results of CaseQ with different base models. The *Drop* measures the percentage of decrease of a metric when time gap size increases.

Experimental Results

Datasets: Movielens, Yelp, Stack Overflow, ATM Maintenance:

1.) Metrics:
   a.) Normalized Discounted Cumulative Gain (NDCG@10)
   b.) Hit Ratio (HR@10)

2.) Findings:
   a.) CaseQ reduces performance drop due to distribution shifts by 47.77% for NDCG and 35.73% for HR compared to standard models.
   b.) CaseQ generalizes better than existing models, showing more robustness against time-based shifts.

Limitations of CaseQ

CaseQ Limitations:

- Fixed Context Types: Assumes a predefined number of contexts
- KL Divergence Regularization: Can lead to over-regularization
- Rigid Hierarchical Structure: Limited adaptability to new unseen contexts
- Inference Complexity: Variational inference adds expensive computational overhead.

Innovations - Meta-Learning

Key Innovation: Meta-Learned Causal Inference

- This takes a more holistic approach by training models on diverse tasks and datasets based on data assumptions.
- Meta-Learning for Fast Adaptation – handles unseen contexts more efficiently

Model Agnostic Meta-Learning (MAML) Optimization:

- Loss function for event prediction
- Training and test sequences for a given context
- Inner update: Adjust model parameters per environment
- Outer update: Train a meta-model that generalizes across environments

Why is this better?

- CaseQ assumes a fixed number of latent contexts making it rigid.
- We dynamically estimate contexts in a few-shot manner, allowing adaptation.
- Does not require any form of KL regularization which avoids over-regularization issues.

Innovations - Causal Bayesian Networks

While CaseQ uses variational inference for context modeling, I will propose a graphical approach using Bayesian networks.

Key Idea:

- Instead of assuming that $P(Y|S)$ is fully observed, we will explicitly model latent causal structures.
- We construct a CBN that represents cause-effect relationships in the data.

Structural Causal Model Formulation:

- We define a causal graph:

  $C \rightarrow S \rightarrow Y$

  $C \rightarrow Y$
- Unlike CaseQ, which estimates $P(Y|S, C)$ indirectly, we can explicitly parameterize it using a Bayesian Network
  - $P(Y|S, C) = \Sigma_c P(Y|S, C = c)P(C = c)$
- However, instead of using a variational approximation for $P(C)$, we learn it dynamically using neural processes.

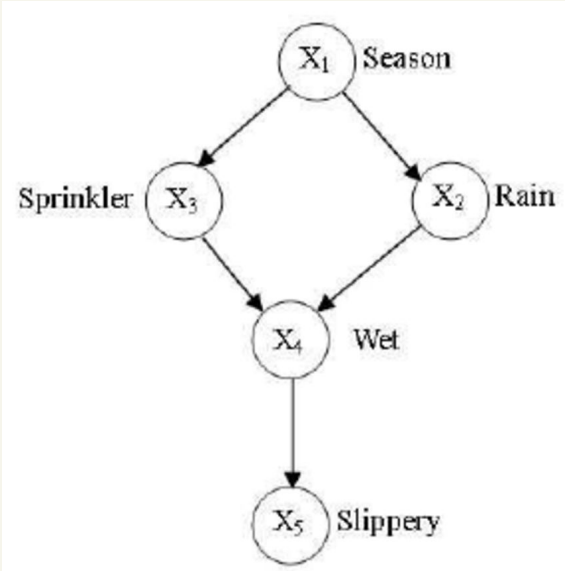
Innovations - Causal Bayesian Networks

Why is this better than CaseQ?

1.) CaseQ's hierarchical structure assumes predefined branches for contexts

2.) Our method infers causal relations directly using Bayesian inference

3.) We avoid hard parameter sharing across different environments which inevitable makes the model more flexible.



Innovations - Neural Processes for Flexible Content Estimation

One limitation of CaseQ is that contexts are treated as discrete variables, which limits flexibility

We then introduce Neural Processes, which model continuous, dynamic context distributions.

Neural Process Formulation:

- A NP learns a distribution over functions, mapping historical event sequences *S* to context embeddings *C*.
  - $C \sim N(\mu(S), \sigma(S))$
- Where $\mu(S), \sigma(S)$ are learned from the training data.

Why is this better?

- This allows the model to dynamically infer contexts as opposed to fixing them to a finite set.
- CaseQ requires sampling from a fixed set of context embeddings, while NPs learn a flexible and continuous representation.

Comparison

Feature	CaseQ	Our Alternative Solution
Context Estimation	Variational Inference (fixed K^D)	Meta-Learned Continuous Contexts
Adaptability	Limited to pre-specified contexts	Fast adaptation to unseen environments
Generalization	Rigid hierarchical structure	Bayesian inference for latent causal discovery
Causal Modeling	Indirect through backdoor adjustment	Explicit causal modeling via Bayesian networks
Flexibility	Requires prior knowledge of number of contexts	Automatically infers latent structure
Optimization	KL Divergence Regularization	Meta-Learning (MAML) for adaptation

Key Takeaways

1.) CaseQ is a strong baseline but has limitations in generalization.

2.) Our solution improves upon CaseQ by:
   a.) Introducing **Meta-Learning** for adaptive context estimation
   b.) Using **Bayesian Networks** to model causal dependencies directly.
   c.) Leveraging **Neural Processes** for flexible, continuous context modeling.

In conclusion, our solution would be more adaptable, scalable, and robust.

Thank you!
Q&A