Enhancement of local embeddings for event sequences data models

Data Science

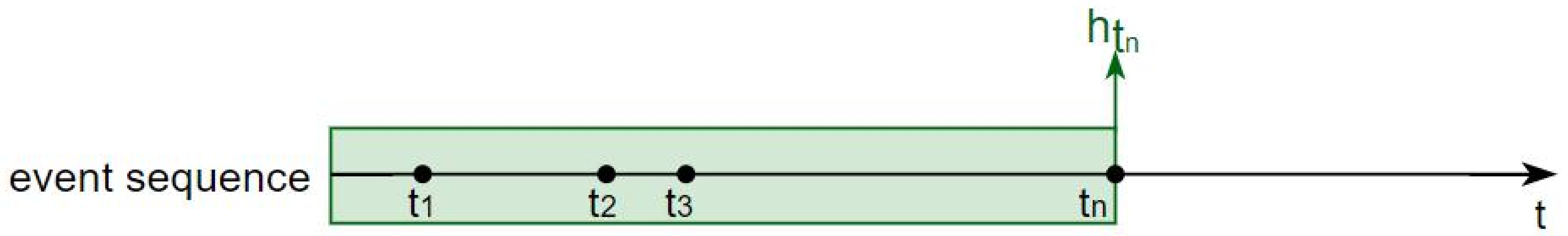
April, 2024

Student: Maria Kovaleva

Research Advisor: Alexey Zaytsev

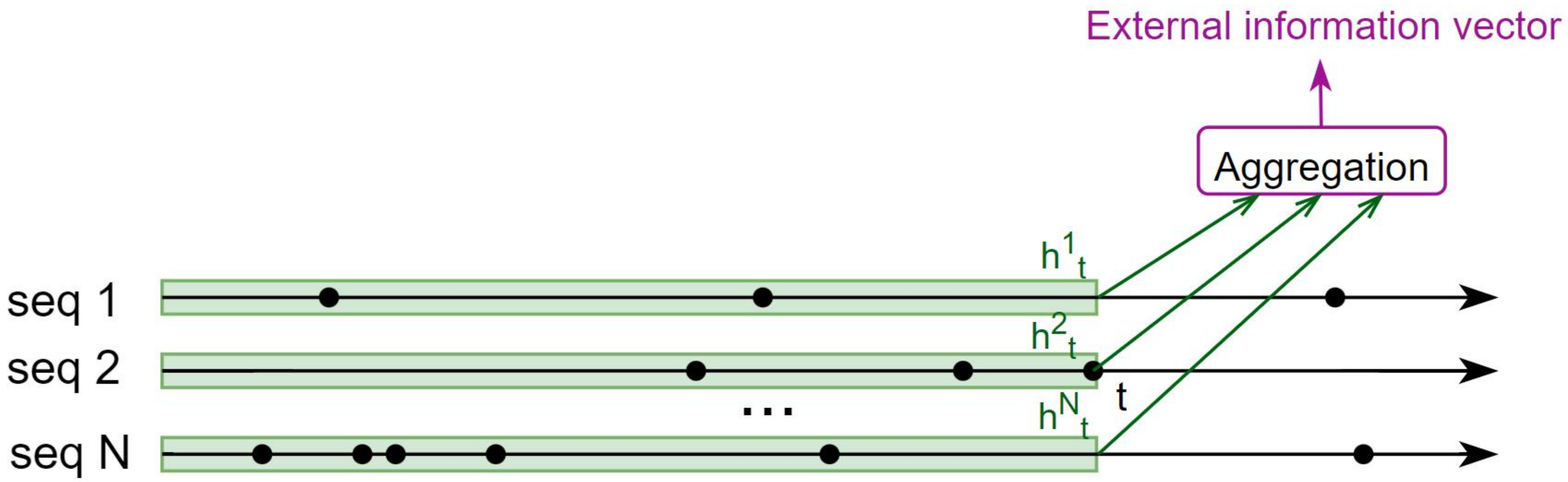
Problem Statement

General problem: building of embeddings $\boldsymbol{h_t}$ for event sequence data



Gaps: self-supervised models for representation learning ignore external information

Idea: the external information is contained in sequences themselves and can be represented as their aggregations



Aim and Objectives

The aim of the work: to enhance the embeddings for event sequences data models using aggregation of external information

Objectives:

1. Study of existing models for event sequence representation learning
2. Development of a method for aggregation
3. Validation of developed methods on bank transaction data

Methods. Basic approach.

Set of n sequences $D = \{S^i\}_{i=1}^{n}$

Each sequence: $S^i = \{(t_j^i, \boldsymbol{Z}_j^i)\}_{j=0}^{T^i}$, where $t_j^i \in [\boldsymbol{0}, \boldsymbol{T^i}]$ - time of event, $\boldsymbol{Z}_j^i \in \boldsymbol{R^d}$ - description of the event

Embeddings construction: $e(S^i) = H^i$

- e is encoder: usually dense NN for event description encoding + recurrent NN
- $H^i = \{(t_j^i, \boldsymbol{h}_j^i)\}_{j=0}^{T^i}$ - embeddings, $\boldsymbol{h}_j^i \in \boldsymbol{R^{d'}}$

Contrastive learning for encoder:
$L_{km} = I_{k=m} \frac{1}{2} d(\boldsymbol{h}^k, \boldsymbol{h}^l)^2 + (1 - I_{k=m}) \frac{1}{2} \max\{0, \rho - d(\boldsymbol{h}^k, \boldsymbol{h}^l)\}^2$, where d is a distance between embeddings and ρ is a hyperparameter

Autoregressive learning for encoder:
A loss consists of the cross-entropy for the categorical features and mean squared error for the continuous features

Methods. Aggregations.

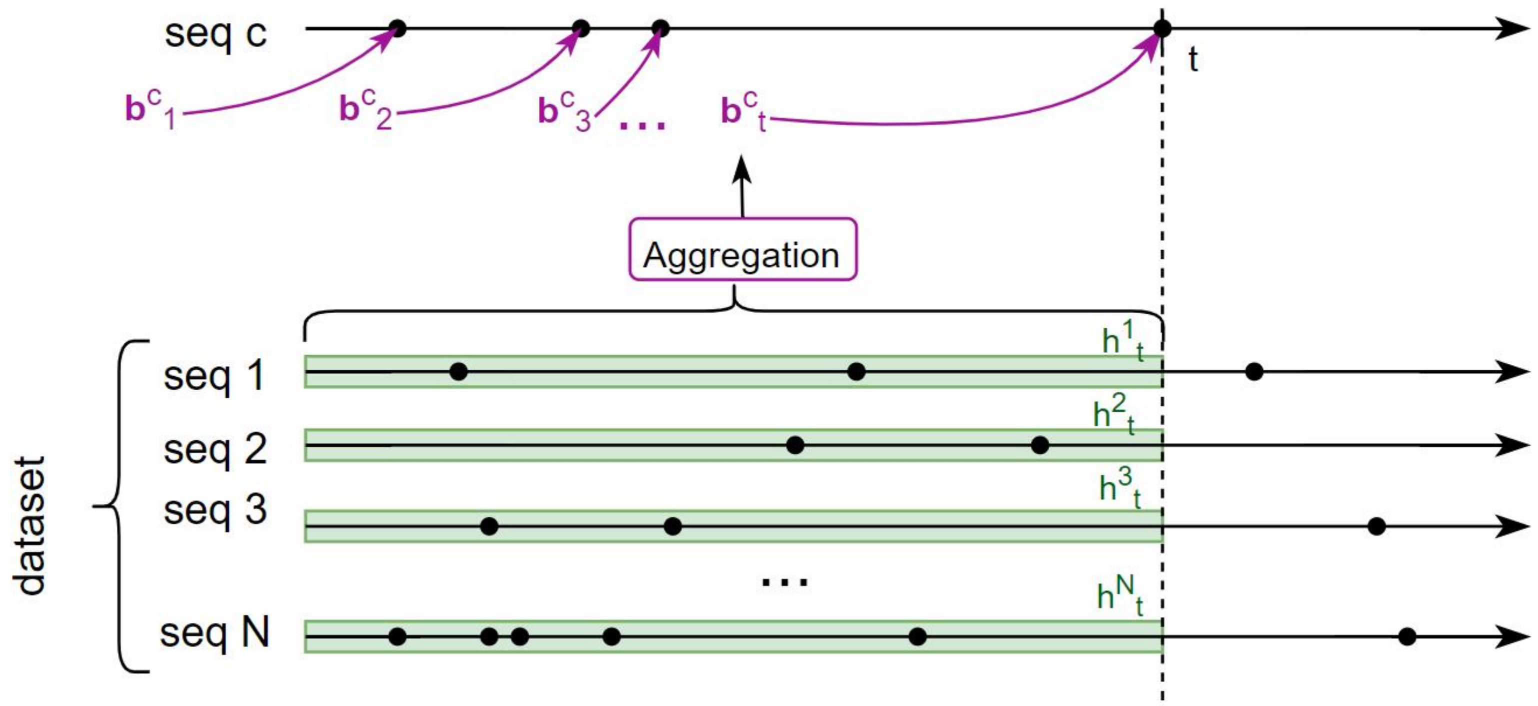
$\boldsymbol{b}_\tau^c$ - vector of external information for user c at the time point τ

Classic:

- **Mean**: $\boldsymbol{b}_\tau^c = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_{t<\tau}^i$
- **Max**: $\boldsymbol{b}_\tau^c = \max(H)$

Attention based:

- **Attention**: $\boldsymbol{b}_\tau^c = H \operatorname{softmax}(H^T \boldsymbol{h}_\tau^c)$
- **Learnable attention**:
$$\boldsymbol{b}_\tau^c = H \operatorname{softmax}(H^T A \boldsymbol{h}_\tau^c)$$
- **Symmetrical attention:**
$$\boldsymbol{b}_\tau^c = H \operatorname{softmax}(H^T S^T S \boldsymbol{h}_\tau^c)$$
- τ
- **Kernel attention:**
$$\boldsymbol{b}_\tau^c = H \operatorname{softmax}(\varphi(H^T)\, \varphi(\boldsymbol{h}_\tau^c))$$



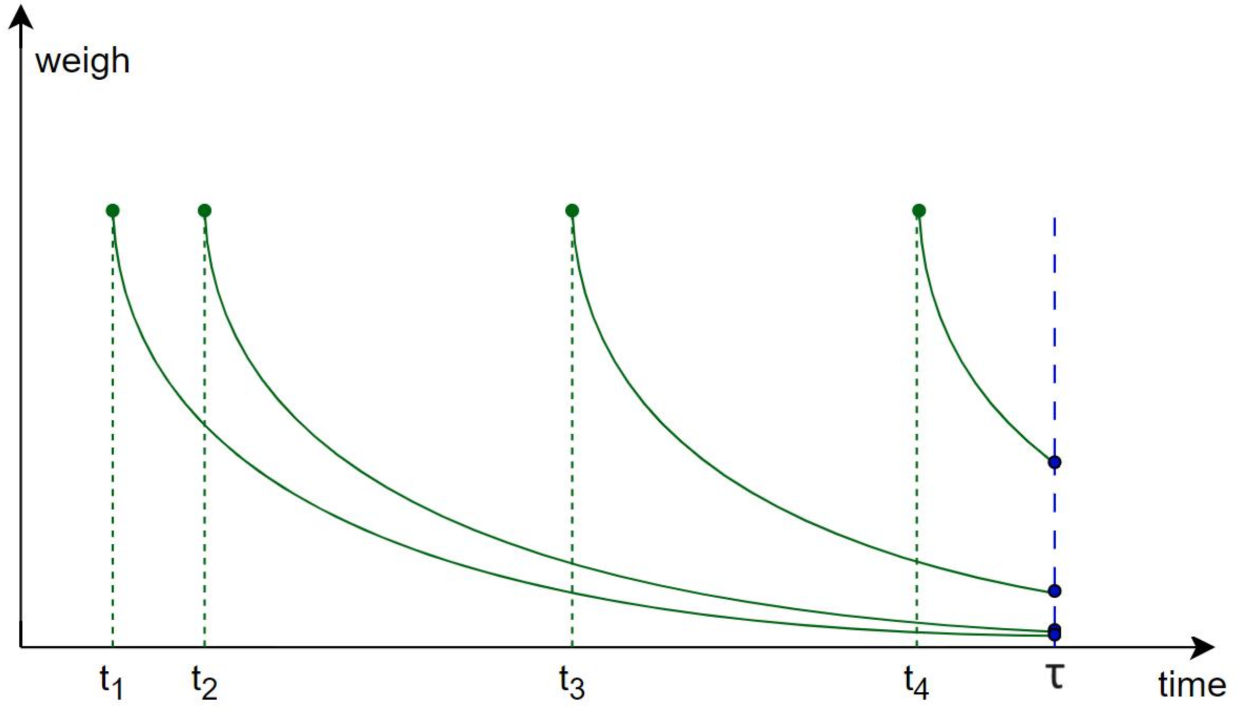
Common pipeline for different aggregations

where $H = [\boldsymbol{h}_{t<\tau}^1, \ldots, \boldsymbol{h}_{t<\tau}^n]$ - matrix with embeddings of all sequences at current time,
A, S - matrices with learnable parameters, φ - learnable transformation (two-layer NN)

Methods. Aggregations.

Aggregations inspired by Hawkes process:

- **Exp Hawkes:** $\boldsymbol{b}_\tau^c = H\exp(-(\tau \cdot \overline{\mathbf{1}} - T))$
- **Exp learnable Hawkes**:
$$\boldsymbol{b}_\tau^c = \varphi_{NN}(\mathrm{concat}(\mathrm{H}, \boldsymbol{h}_\tau^c))\exp(-(\tau \cdot \overline{\mathbf{1}} - T))$$
- **Attention Hawkes**:
$$\boldsymbol{b}_\tau^c = X\,\mathrm{softmax}(H^T\,\boldsymbol{h}_\tau^c)\exp(-(\tau \cdot \overline{\mathbf{1}} - T))$$



where $H = [\boldsymbol{h}_{t<\tau}^1, \ldots, \boldsymbol{h}_{t<\tau}^n]$ - matrix with embeddings of all sequences at current time,
$T = [t_{t<\tau}^1, \ldots, t_{t<\tau}^{\boldsymbol{n}}]$ - vector of last event times for all embeddings $\boldsymbol{h}_{t<\tau}^1, \ldots, \boldsymbol{h}_{t<\tau}^n$,
φ_{NN} -learnable transformation (two-layer NN)

Methods. Validation and Datasets.

Dataset of bank transactions:

- Each sequence S^i: one bank client transactions
- Each event $\boldsymbol{Z}_j^i$: transaction (merchant category code + amount)
- Target: whether the client left the bank

Validation:

- **Global** – the model inference on the whole sequence to check its global patterns.
  - binary downstream task
  - one output vector for a boosting model
  - ROC AUC
- **Local** – the model inference on sliding windows to check local properties.
  a. next event type prediction
  b. local target prediction
  - MLP head for prediction
  - ROC AUC

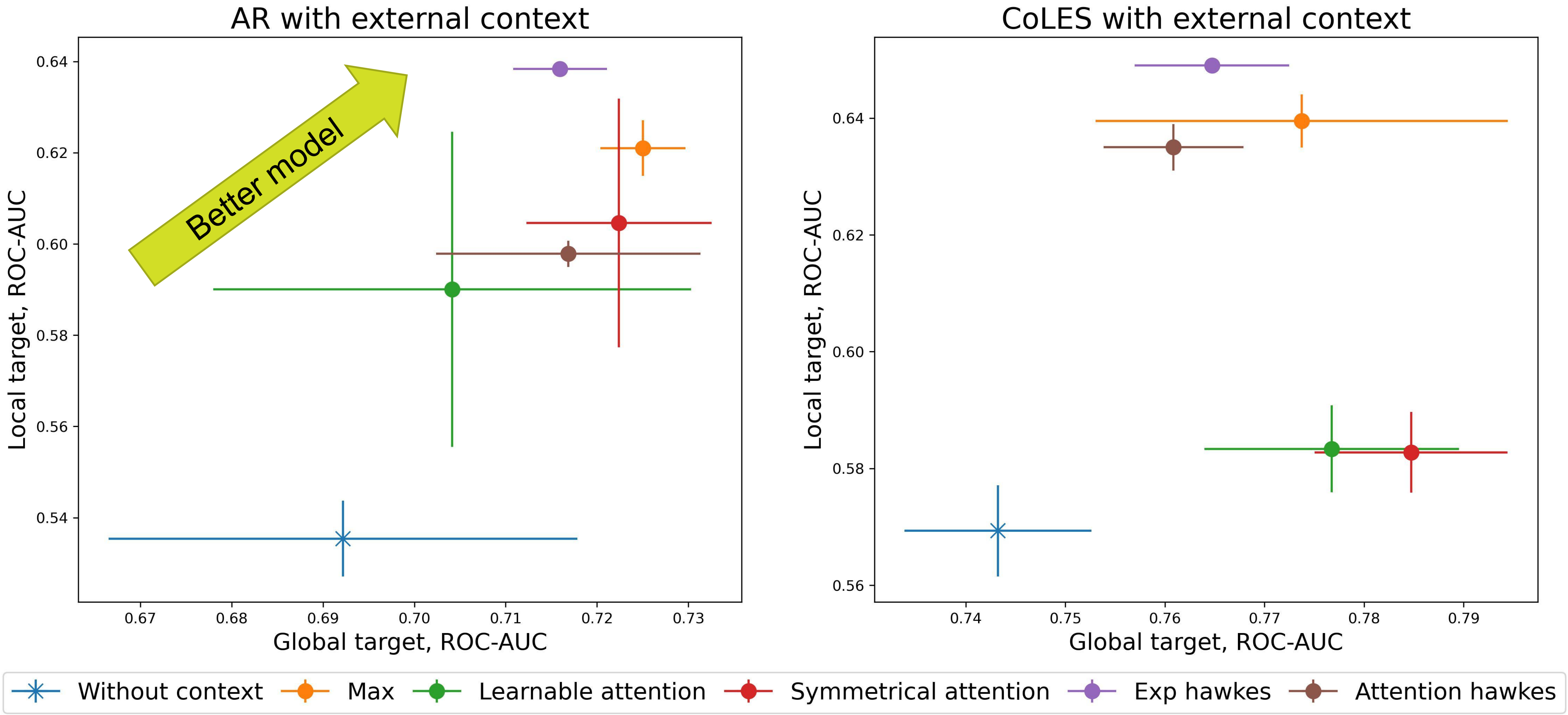
Results

(blue)	1st results
(green)	2nd results
(yellow)	3rd results

Global target	Contrastive learning		Autoregressive learning	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Without context	0.743 ± 0.009	0.792 ± 0.014	0.692 ± 0.025	0.734 ± 0.032
Mean	0.773 ± 0.004	0.828 ± 0.003	0.722 ± 0.007	0.776 ± 0.005
Max	0.774 ± 0.021	0.818 ± 0.032	0.725 ± 0.005	0.777 ± 0.002
Attention	0.760 ± 0.014	0.808 ± 0.017	0.696 ± 0.014	0.744 ± 0.017
Learn. attention	0.777 ± 0.013	0.830 ± 0.013	0.704 ± 0.026	0.751 ± 0.020
Sym. attention	0.785 ± 0.010	0.835 ± 0.005	0.722 ± 0.010	0.769 ± 0.004
Kernel attention	0.775 ± 0.003	0.824 ± 0.002	0.709 ± 0.019	0.760 ± 0.003
Exp Hawkes	0.765 ± 0.008	0.814 ± 0.009	0.716 ± 0.005	0.767 ± 0.013
Exp learn. Hawkes	0.764 ± 0.008	0.812 ± 0.008	0.714 ± 0.025	0.758 ± 0.020
Attention Hawkes	0.761 ± 0.007	0.796 ± 0.009	0.717 ± 0.014	0.751 ± 0.023

Local target	Contrastive learning		Autoregressive learning	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Without context	0.569 ± 0.008	0.321 ± 0.003	0.535 ± 0.008	0.299 ± 0.011
Mean	0.592 ± 0.005	0.342 ± 0.005	0.543 ± 0.006	0.312 ± 0.006
Max	0.640 ± 0.005	0.400 ± 0.006	0.621 ± 0.006	0.256 ± 0.008
Attention	0.600 ± 0.009	0.348 ± 0.010	0.534 ± 0.016	0.301 ± 0.007
Learn. attention	0.583 ± 0.007	0.330 ± 0.008	0.590 ± 0.035	0.338 ± 0.025
Sym. attention	0.583 ± 0.007	0.329 ± 0.007	0.605 ± 0.027	0.350 ± 0.020
Kernel attention	0.582 ± 0.007	0.329 ± 0.007	0.572 ± 0.021	0.330 ± 0.023
Exp Hawkes	0.649 ± 0.000	0.366 ± 0.003	0.638 ± 0.001	0.351 ± 0.001
Exp learn. Hawkes	0.581 ± 0.012	0.322 ± 0.013	0.539 ± 0.034	0.293 ± 0.025
Attention Hawkes	0.635 ± 0.004	0.359 ± 0.005	0.598 ± 0.003	0.331 ± 0.001

Results



- **External information** enhances metrics
- **Exp Hawkes** method is the best for the local task
- **Classical** methods are the best for the global task
- **Attention-based** aggregations help to highlight local patterns in sequences

Conclusions

- External information addition improves the embeddings of event sequences
- Aggregation of all sequences at the current time point can represent the external information
- We propose new methods of aggregation of external information
- Experiments with bank transactions data show the impact of the proposed methods

Outcomes

Paper: Bazarova, A.*, Kovaleva, M.*, Kuleshov, I.*, Romanenkova, E.*, Stepikin, A.*, Yugay, A.*, Mollaev, D., Kireev, I., Savchenko, A., and Zaytsev, A*. Universal representations for financial transactional data: embracing local, global, and external contexts. arXiv preprint arXiv:2404.02047 (2024)

Contribution: The article contains an overview of various methods for obtaining embeddings for transactional data. I was responsible for the part dedicated to the aggregation of external information.

*equal contribution

Acknowledgements

Alexey Zaytsev
Evgenia Romanenkova
Ilya Kulesov
Alexander Yugai
Alexandra Bazarova
Alexander Stepikin

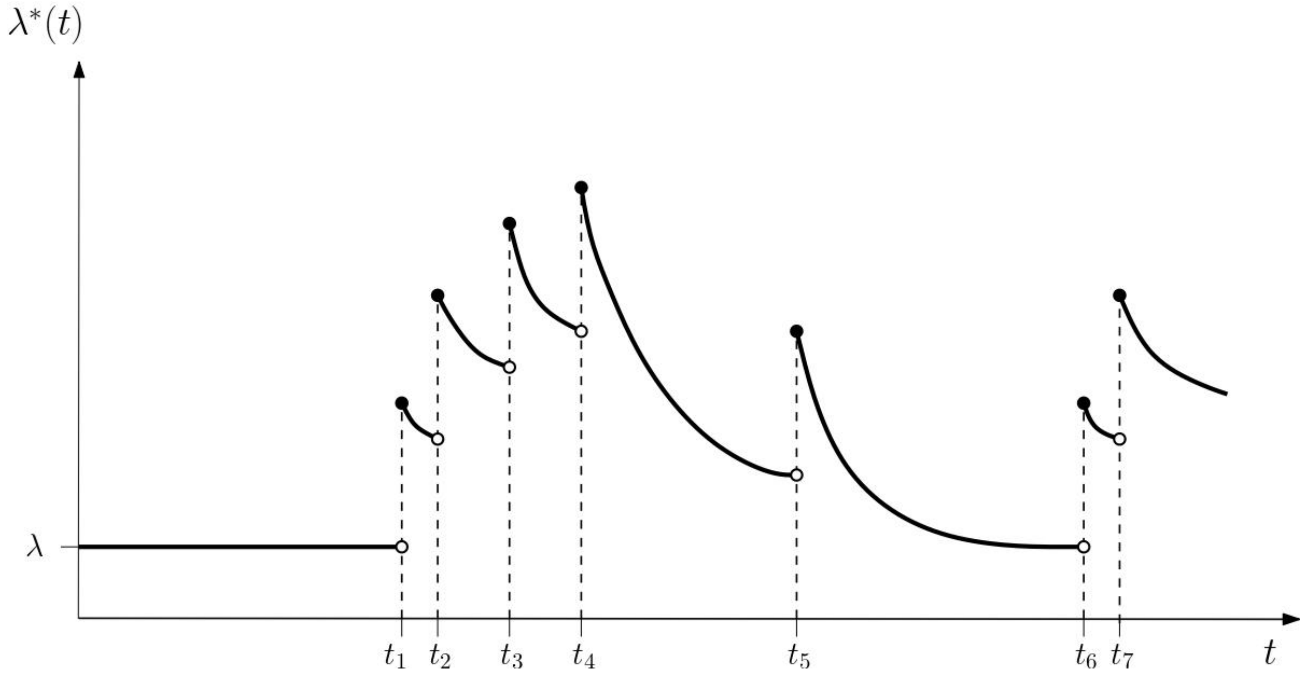
Thx!

Hawkes process

Hawkes process is a self-exciting temporal point process whose conditional intensity function $\lambda = \lambda(\mathrm{t})$ is defined to be:

$$\lambda(\mathrm{t}) = \mu(\mathrm{t}) + \sum_{i:\tau_i<t} \nu(t-\tau_i)$$

where $\mu(\mathrm{t})$ is the background rate of the process, τ_i are the points in time occurring prior to time t, and where ν is a function which defines the density of the process.



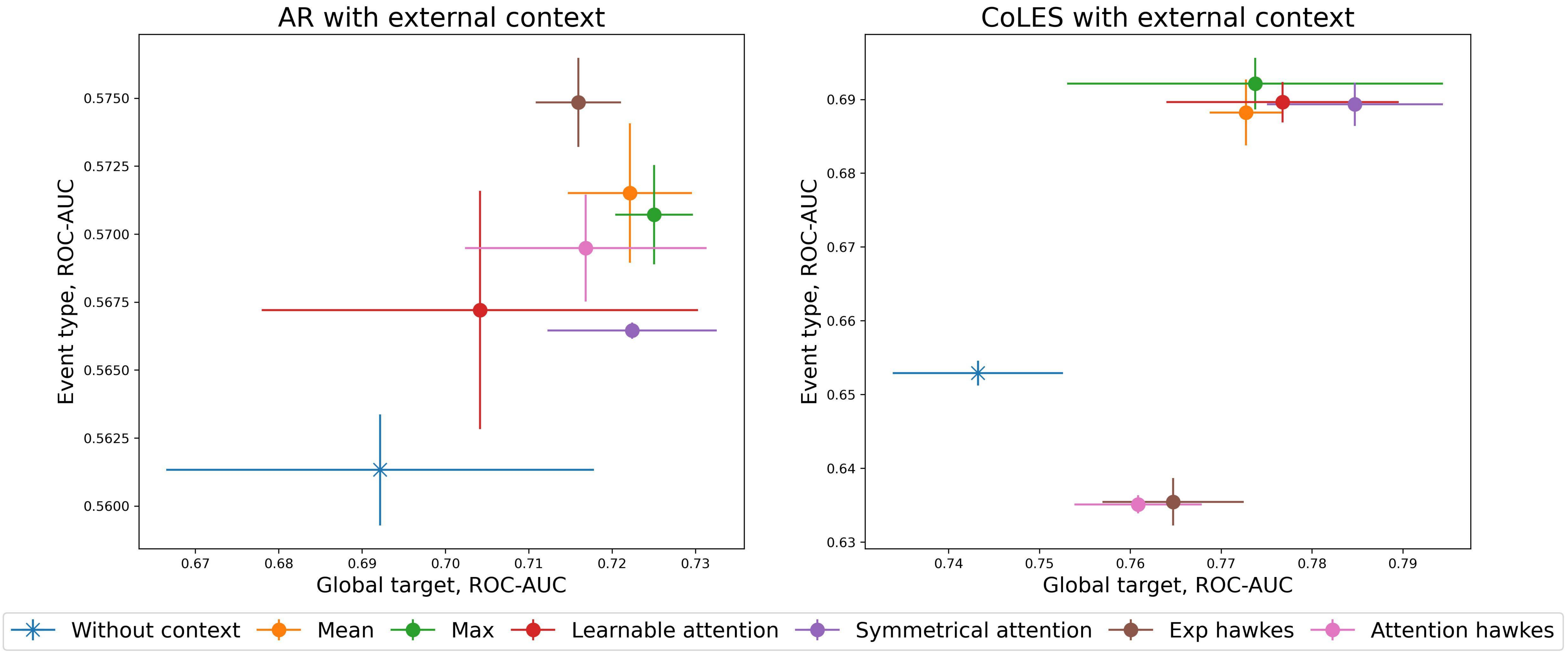
Laub, Patrick J., Thomas Taimre, and Philip K. Pollett. "Hawkes processes." arXiv preprint arXiv:1507.02822 (2015).

Datasets details

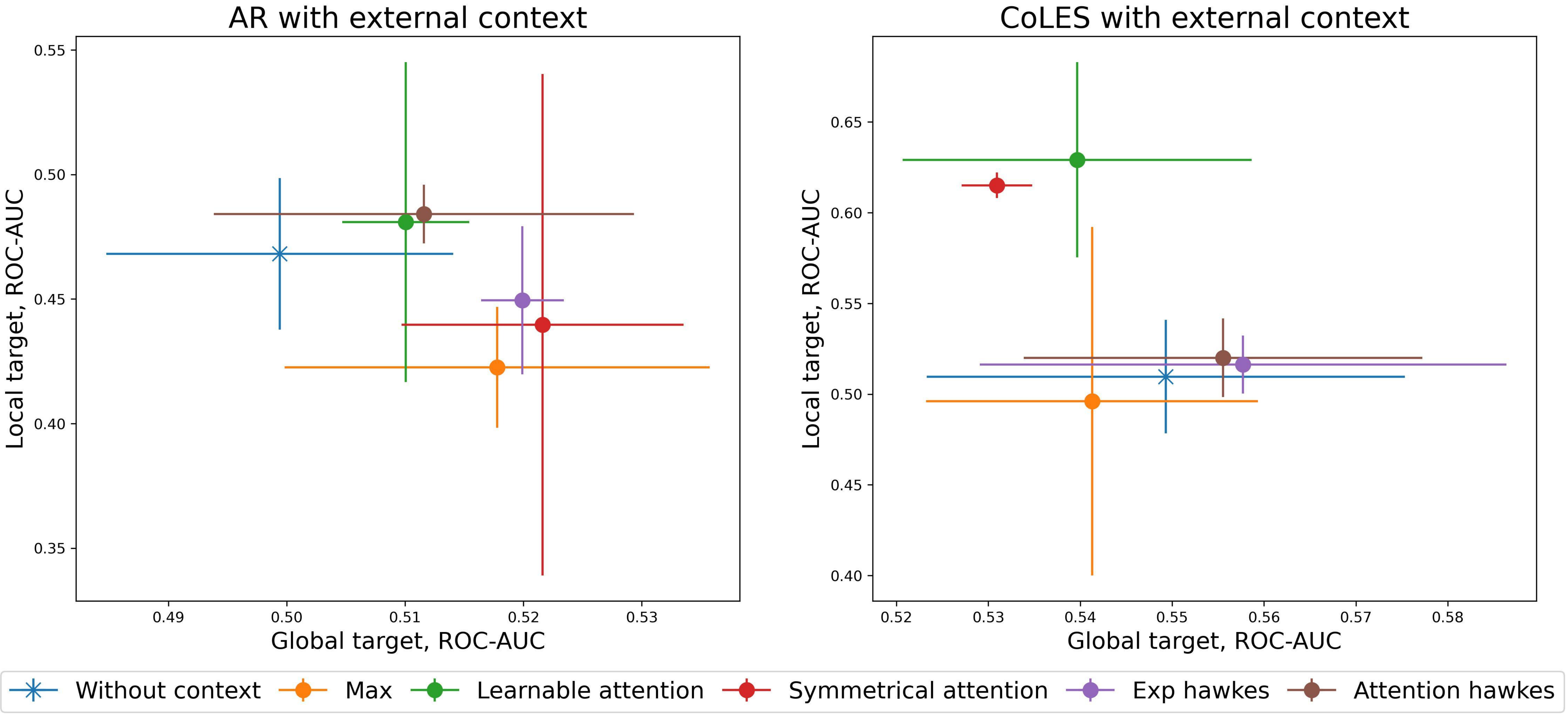
- **Churn:**
  - ~500k transactions, 5000 users
  - target flag: a binary label, whether the client has stopped doing business with this bank
  - 10 fields

- **Default:**
  - ~2.1m transactions, 7080 users
  - 5 fields
  - target: binary label, whether a client will default
  - all transaction lengths are equal to 300

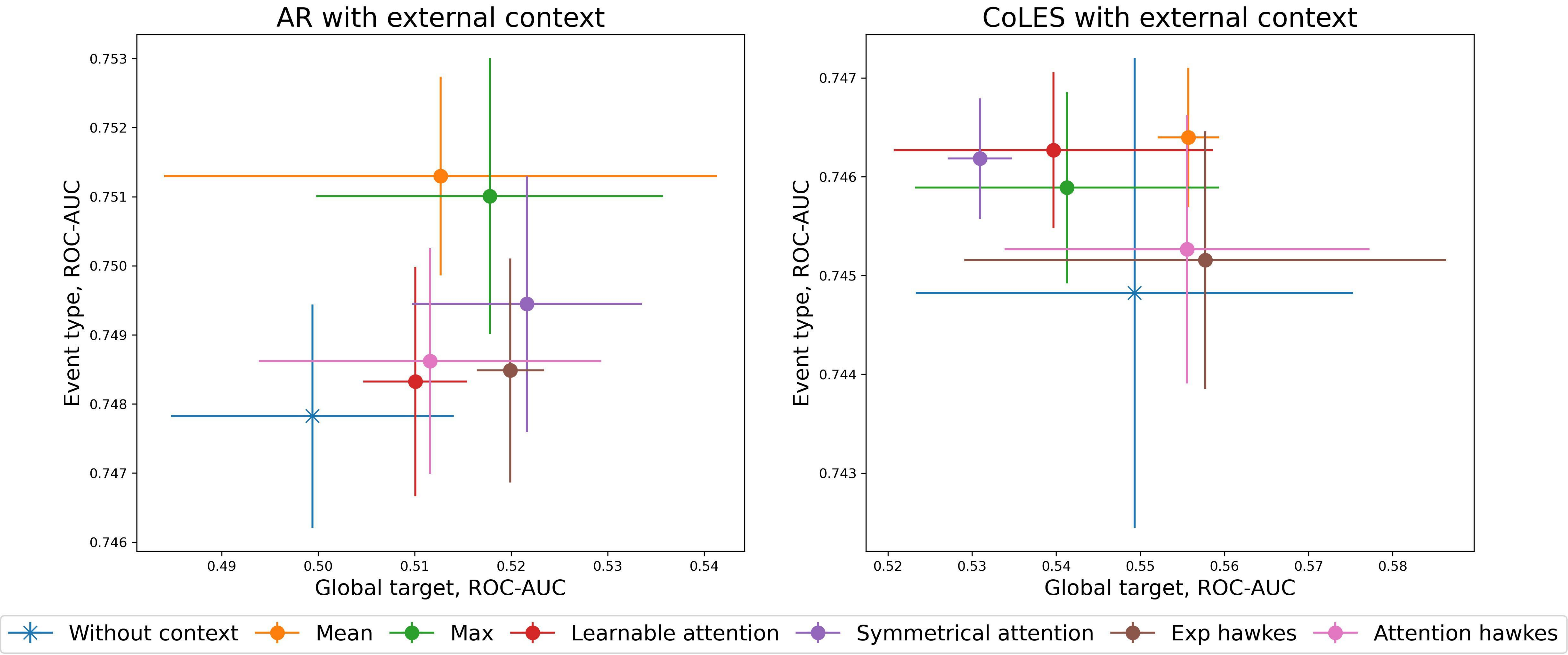
Additional results. Churn.



Additional results. Default.



Additional results. Default.



Results. Default.

Global target						
	Contrastive learning			Autoregressive learning		
	ROC-AUC	PR-AUC	Accuracy	ROC-AUC	PR-AUC	Accuracy
Without context	0.743 ± 0.009	0.792 ± 0.014	0.689 ± 0.005	0.692 ± 0.025	0.734 ± 0.032	0.657 ± 0.013
Mean	0.773 ± 0.004	0.828 ± 0.003	0.715 ± 0.010	0.722 ± 0.007	0.776 ± 0.005	0.653 ± 0.008
Max	0.774 ± 0.021	0.818 ± 0.032	0.701 ± 0.004	0.725 ± 0.005	0.777 ± 0.002	0.664 ± 0.012
Attention	0.760 ± 0.014	0.808 ± 0.017	0.691 ± 0.010	0.696 ± 0.014	0.744 ± 0.017	0.644 ± 0.022
Learn. attention	0.777 ± 0.013	0.830 ± 0.013	0.699 ± 0.006	0.704 ± 0.026	0.751 ± 0.020	0.655 ± 0.009
Sym. attention	0.785 ± 0.010	0.835 ± 0.005	0.703 ± 0.017	0.722 ± 0.010	0.769 ± 0.004	0.671 ± 0.009
Kernel attention	0.775 ± 0.003	0.824 ± 0.002	0.693 ± 0.009	0.709 ± 0.019	0.760 ± 0.003	0.655 ± 0.014
Exp hawkes	0.765 ± 0.008	0.814 ± 0.009	0.699 ± 0.012	0.716 ± 0.005	0.767 ± 0.013	0.661 ± 0.011
Exp learn. hawkes	0.764 ± 0.008	0.812 ± 0.008	0.703 ± 0.006	0.714 ± 0.025	0.758 ± 0.020	0.665 ± 0.024
Attention hawkes	0.761 ± 0.007	0.796 ± 0.009	0.702 ± 0.007	0.717 ± 0.014	0.751 ± 0.023	0.667 ± 0.006

	1st results
	2nd results
	3rd results

	Local target				Event type					
	Contrastive learning		Autoregressive learning		Contrastive learning			Autoregressive learning		
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	Accuracy	ROC-AUC	PR-AUC	Accuracy
Without context	0.569 ± 0.008	0.321 ± 0.003	0.535 ± 0.008	0.299 ± 0.011	0.653 ± 0.002	0.168 ± 0.001	0.239 ± 0.003	0.561 ± 0.002	0.111 ± 0.002	0.237 ± 0.002
Mean	0.592 ± 0.005	0.342 ± 0.005	0.543 ± 0.006	0.312 ± 0.006	0.688 ± 0.004	0.193 ± 0.003	0.242 ± 0.001	0.572 ± 0.003	0.119 ± 0.003	0.238 ± 0.004
Max	0.640 ± 0.005	0.400 ± 0.006	0.621 ± 0.006	0.256 ± 0.008	0.692 ± 0.003	0.192 ± 0.001	0.239 ± 0.002	0.570 ± 0.002	0.115 ± 0.001	0.231 ± 0.002
Attention	0.600 ± 0.009	0.348 ± 0.010	0.534 ± 0.016	0.301 ± 0.007	0.691 ± 0.004	0.194 ± 0.004	0.243 ± 0.001	0.571 ± 0.004	0.117 ± 0.002	0.236 ± 0.005
Learn. attention	0.583 ± 0.007	0.330 ± 0.008	0.590 ± 0.035	0.338 ± 0.025	0.690 ± 0.003	0.194 ± 0.004	0.241 ± 0.001	0.567 ± 0.004	0.116 ± 0.001	0.238 ± 0.003
Sym. attention	0.583 ± 0.007	0.329 ± 0.007	0.605 ± 0.027	0.350 ± 0.020	0.689 ± 0.003	0.193 ± 0.004	0.241± 0.0004	0.566 ± 0.000	0.115 ± 0.002	0.238 ± 0.003
Kernel attention	0.582 ± 0.007	0.329 ± 0.007	0.572 ± 0.021	0.330 ± 0.023	0.689 ± 0.003	0.193 ± 0.004	0.241 ± 0.000	0.566 ± 0.004	0.115 ± 0.001	0.238 ± 0.002
Exp hawkes	0.649 ± 0.000	0.366 ± 0.003	0.638 ± 0.001	0.351 ± 0.001	0.635 ± 0.003	0.161 ± 0.002	0.244 ± 0.002	0.575 ± 0.002	0.122 ± 0.001	0.244 ± 0.001
Exp learn. hawkes	0.581 ± 0.012	0.322 ± 0.013	0.539 ± 0.034	0.293 ± 0.025	0.613 ± 0.012	0.153 ± 0.007	0.257 ± 0.011	0.549 ± 0.001	0.113 ± 0.002	0.241 ± 0.005
Attention hawkes	0.635 ± 0.004	0.359 ± 0.005	0.598 ± 0.003	0.331 ± 0.001	0.635 ± 0.001	0.159 ± 0.002	0.246 ± 0.001	0.569 ± 0.002	0.118 ± 0.001	0.241 ± 0.003