LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
LMU MUNICH SCHOOL OF MANAGEMENT
INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Causal Transformer for Estimating Counterfactual Outcomes

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

LMU Munich, Munich, Germany

ICML 2022, Short Presentation

Introduction: Estimating counterfactual outcomes over time

Why this is important?

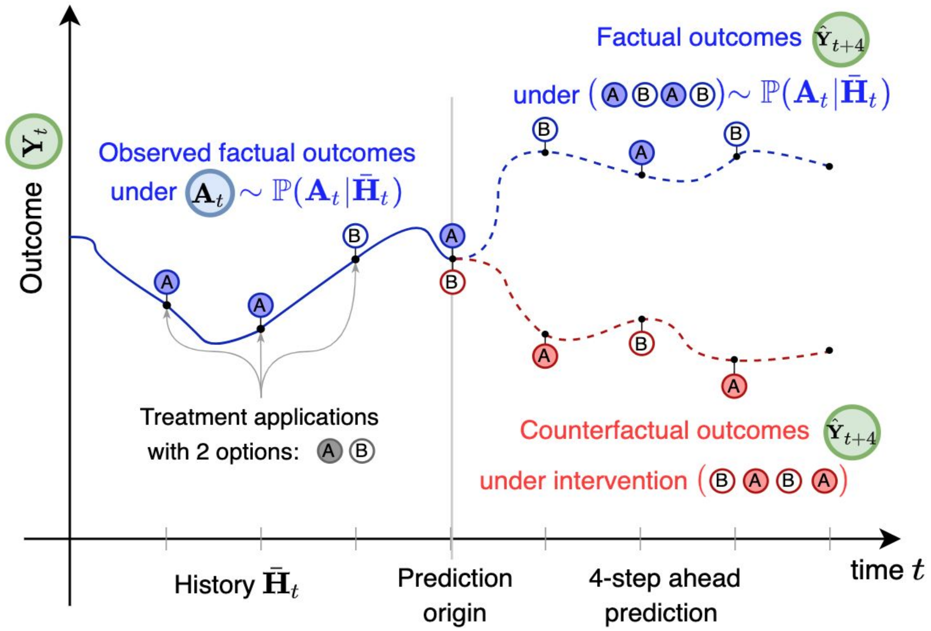
- Counterfactual prediction allows to answer **individualized** “what if” questions: what will happen to the patient, if I apply alternative sequence of treatments, **counterfactual**[1] to a standard treatment policy
- Growing opportunity to employ **observational data**:
  - randomized controlled trials (RCTs) are costly and/or unethical
  - abundance of large-scale observational data, e.g., electronic health records

Problem formulation

Given observational dataset of:

- $\mathbf{X}_t$ time-varying covariates (e.g., blood pressure)
- $\mathbf{V}$ static covariates (e.g., age)
- $\mathbf{A}_t$ treatments (e.g., ventilation)
- $\mathbf{Y}_t$ (factual[2]) outcomes (e.g., respiratory frequency)

we want to estimate **counterfactual outcomes over time** starting from prediction origin for a given sequence of treatment interventions


Factual outcomes $\hat{\mathbf{Y}}_{t+4}$
under (A B A B) $\sim \mathbb{P}(\mathbf{A}_t \mid \bar{\mathbf{H}}_t)$
Observed factual outcomes
under $\mathbf{A}_t \sim \mathbb{P}(\mathbf{A}_t \mid \bar{\mathbf{H}}_t)$
Outcome $\mathbf{Y}_t$
Treatment applications
with 2 options: A B
Counterfactual outcomes $\hat{\mathbf{Y}}_{t+4}$
under intervention (B A B A)
History $\bar{\mathbf{H}}_t$
Prediction origin
4-step ahead prediction
time t

[1] Here, potential outcomes are meant, which correspond to the interventional level of valuation in Pearl’s Hierarchy and the Foundations of Causal Inference

[2] Factual outcomes are observed under standard treatment policy

Introduction: Task complexity – Assumptions – Related methods

Why estimation is hard?

- Counterfactual outcomes are never directly observed in a real world
- Observed history grows with time
- Traditional machine learning is biased in the presence of time-varying confounding[1]

Identifiability assumptions

- **Consistency**. If $\bar{\mathbf{A}}_t = \bar{\mathbf{a}}_t$ is a given sequence of treatments for some patient, then $\mathbf{Y}_{t+1}[\bar{\mathbf{a}}_t] = \mathbf{Y}_{t+1}.$
- **Sequential Overlap.** There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the previous history: $0 < \mathbb{P}(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{H}}_t = \bar{\mathbf{h}}_t) < 1$
- **Sequential Ignorability.** Current treatment is independent of the potential outcome, conditioning on the observed history $\mathbf{A}_t \perp\!\!\!\perp \mathbf{Y}_{t+1}[\mathbf{a}_t] \mid \bar{\mathbf{H}}_t$

Related methods

- **Marginal Structural Models (MSMs)** (Robins et al., 2000; Hernan et al., 2001): only linear modelling
- **Recurrent Marginal Structural Networks** (RMSNs) (Lim et al., 2018): several LSTM networks for inverse probability of treatment weights (IPTW) and prediction
- **Counterfactual Recurrent Network** (CRN) (Bica et al., 2020): encoder-decoder LSTMS with adversarial learning of treatment invariant representations
- **G-Net** (Li et al., 2021): G-computation on top of LSTM

[1] Time-varying confounding stands for a non-randomized treatment assignment, which depends on time-varying covariates, previous treatments and previous outcomes

Introduction: Research gap – Our contributions

Research gap

- Current state-of-the-art methods are built on top of long short-term memory (LSTM), thus rendering inferences for complex, long-range dependencies challenging

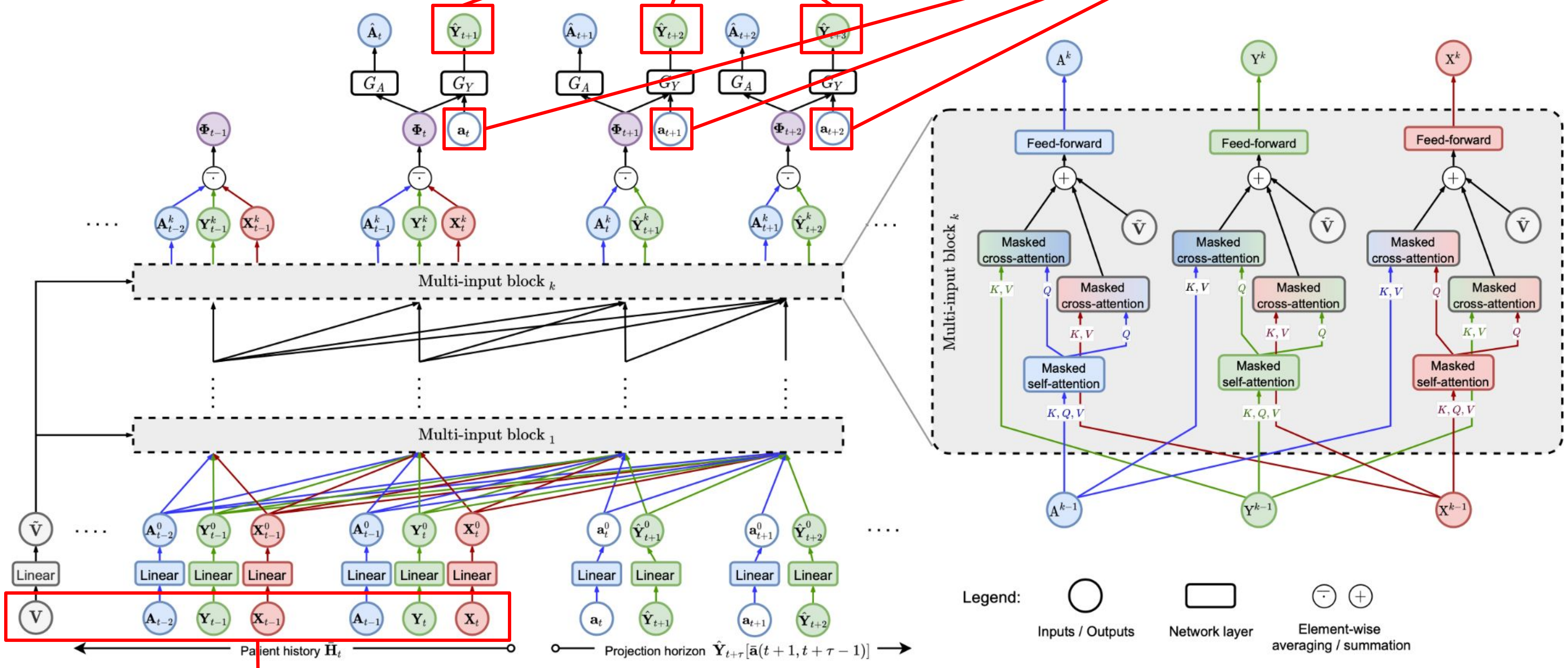
Our contributions

Causal Transformer (CT) is an end-to-end model, first tailoring of transformers to a counterfactual prediction task over time:

- CT captures **complex, long-range dependencies** between time-varying covariates, treatments and outcomes
- CT employs a novel **counterfactual domain confusion (CDC) loss** to address a time-varying confounding
- CT achieves **state-of-the-art performance** on synthetic, semi-synthetic & real benchmarks

Causal Transformer: Novel architecture

2. Output – predicted outcomes under a sequence of interventions

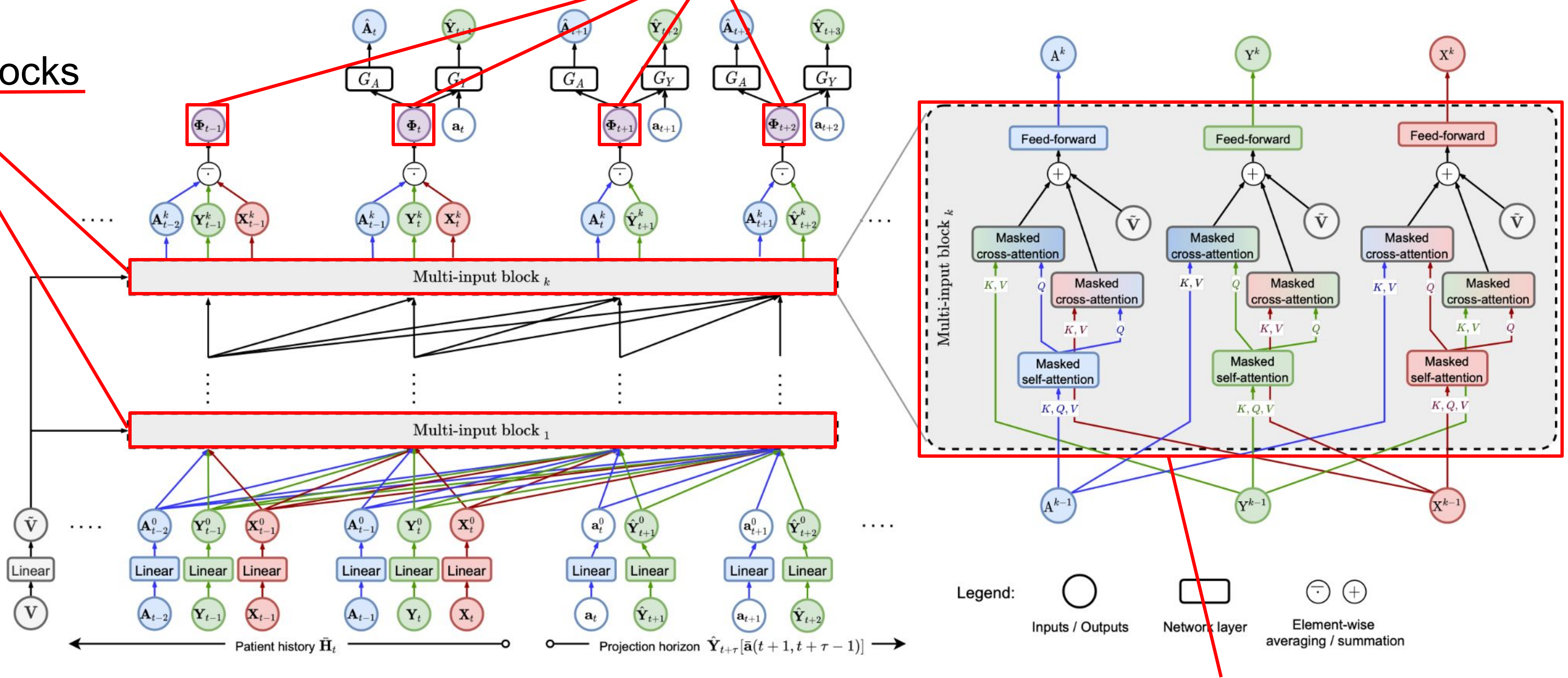


1. Input – observed patient history

Causal Transformer: Novel architecture

3. Inputs are transformed with a stack of multi-input blocks

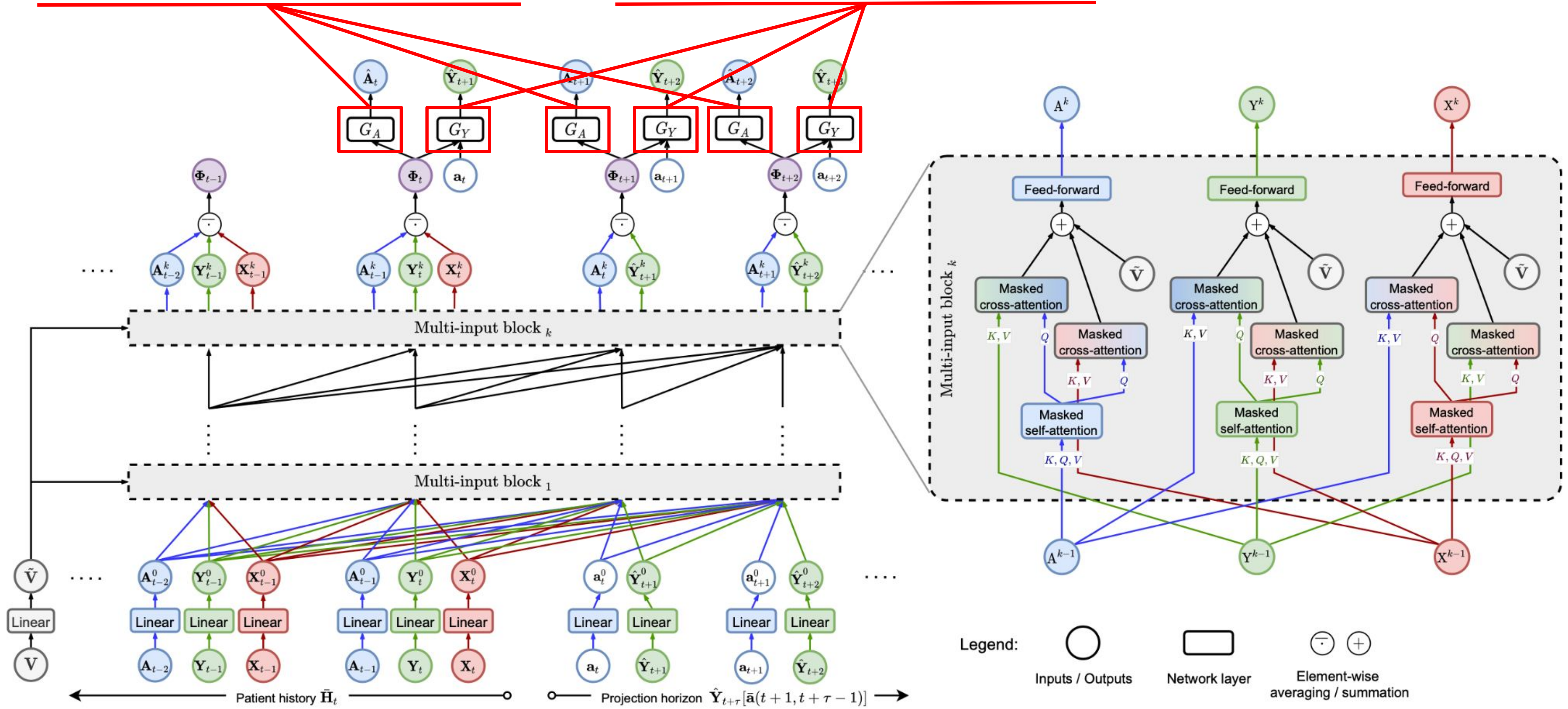
4. Outputs of the last block are averaged and form balanced representations



5. Each block is equipped with self-attention, cross-attention and feed-forward layers

Causal Transformer: Novel architecture

6. We place treatment classifier network and outcome prediction network on top of balanced representations



7. Both treatment classifier and outcome prediction networks are used for the novel counterfactual domain confusion loss (CDC) loss

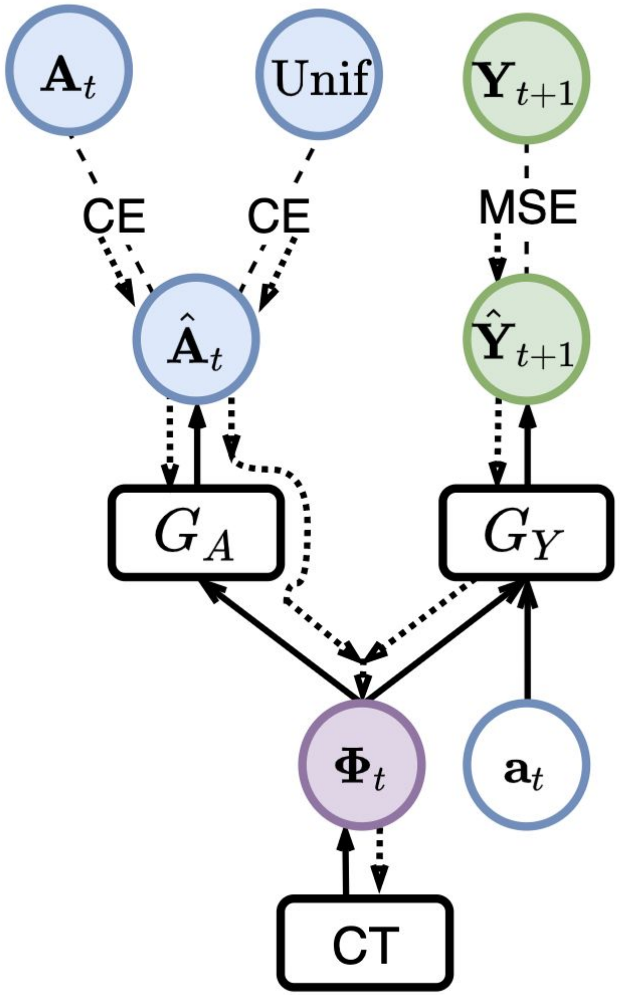
Causal Transformer: Counterfactual domain confusion (CDC) loss

- Idea stems from the unsupervised domain adaptation[1]
- CDC is an adversarial objective, which aims to

  (a) make **balanced representations** Φ_t non-predictive of the current treatment:

  - minimizing cross-entropy of current treatment wrt. $\boxed{G_A}$
  - minimizing cross-entropy between uniform treatment and output of treatment classifier network wrt. $\boxed{\text{CT}}$

  (b) at same time, make them **predictive of the outcome** wrt. $\boxed{\text{CT}}$ and $\boxed{G_Y}$ by minimizing factual MSE
- Adversarial learning is stabilized with exponential moving average (EMA of model weights



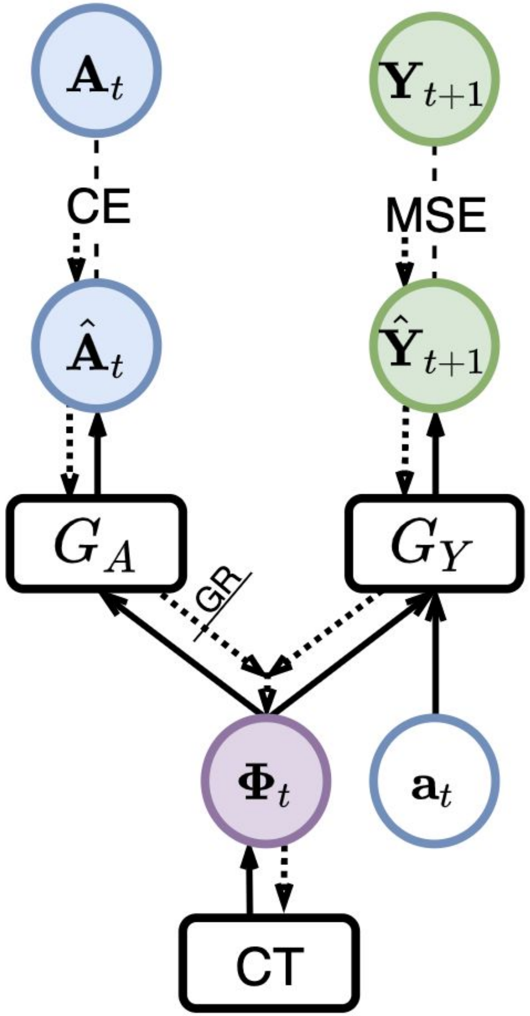
[1] Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." Proceedings of the IEEE international conference on computer vision (2015)

Causal Transformer: Theoretical insights

- Previously proposed **Gradient reversal**[1] (CRN, Bica et al., 2020) extends in two ways:
  - if badly chosen hyperparameter -> representation may be predictive of opposite treatment
  - gradients could vanish, if treatment classifier network learns too fast, gradients vanish
- We prove a theorem, similar to (CRN, Bica et al., 2020): finding a solution to an adversarial objective of CDC loss renders distributions of representations conditional on each treatment **equal** (= balanced)
- In our case, we minimize a reversed KL-divergence:

CDC loss (our paper)	Gradient reversal (CRN, Bica et al., 2020)
Minimizing $\sum_{j=1}^{K} KL\left(\frac{1}{K}\sum_{i=1}^{K} P_i^{\Phi}(x') \middle\| P_j^{\Phi}(x')\right)$	Minimizing $\sum_{j=1}^{K} KL\left(P_j^{\Phi}(x') \middle\| \frac{1}{K}\sum_{i=1}^{K} P_i^{\Phi}(x')\right)$

where $P_j^{\Phi}(x')$ is a distribution of representation conditional on treatment j



[1] Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." International conference on machine learning. PMLR, 2015

Experiments: Datasets – Baselines – Results

Datasets

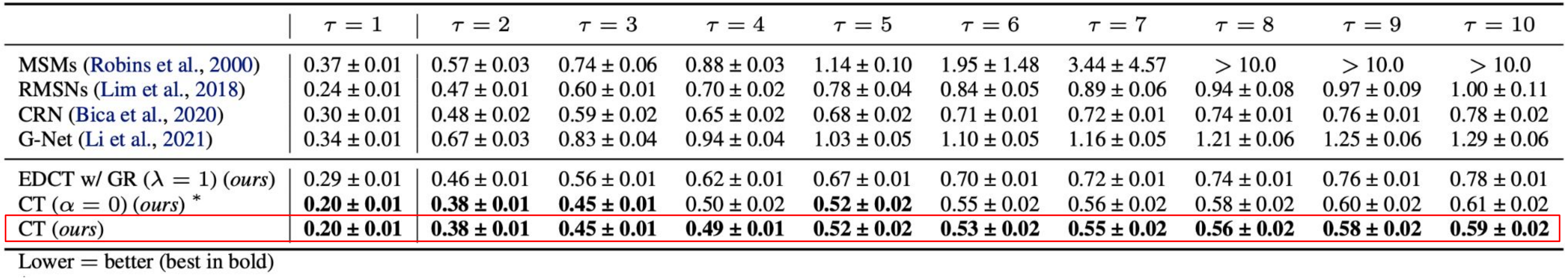
- We evaluate CT based on synthetic, (self-designed) semi-synthetic and real-world (MIMIC-III) datasets
- Only synthetic and semi-synthetic data have ground-truth counterfactuals; real-world evaluation is a proof of concept
- We compared root-mean-squared error (RMSE) of one and multiple-step-ahead predictions

Baselines

- Marginal Structural Models (MSMs) (Robins et al., 2000; Hernan et al., 2001)
- Recurrent Marginal Structural Networks (RMSNs) (Lim et al., 2018)
- Counterfactual Recurrent Network (CRN) (Bica et al., 2020)
- G-Net (Li et al., 2021)

Results

CT achieves **superior performance** over current baselines for benchmarks with long-range dependencies and long prediction horizons, e.g., for semi-synthetic benchmark:



	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
MSMs (Robins et al., 2000)	0.37 ± 0.01	0.57 ± 0.03	0.74 ± 0.06	0.88 ± 0.03	1.14 ± 0.10	1.95 ± 1.48	3.44 ± 4.57	> 10.0	> 10.0	> 10.0
RMSNs (Lim et al., 2018)	0.24 ± 0.01	0.47 ± 0.01	0.60 ± 0.01	0.70 ± 0.02	0.78 ± 0.04	0.84 ± 0.05	0.89 ± 0.06	0.94 ± 0.08	0.97 ± 0.09	1.00 ± 0.11
CRN (Bica et al., 2020)	0.30 ± 0.01	0.48 ± 0.02	0.59 ± 0.02	0.65 ± 0.02	0.68 ± 0.02	0.71 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.02
G-Net (Li et al., 2021)	0.34 ± 0.01	0.67 ± 0.03	0.83 ± 0.04	0.94 ± 0.04	1.03 ± 0.05	1.10 ± 0.05	1.16 ± 0.05	1.21 ± 0.06	1.25 ± 0.06	1.29 ± 0.06
EDCT w/ GR ($\lambda = 1$) (*ours*)	0.29 ± 0.01	0.46 ± 0.01	0.56 ± 0.01	0.62 ± 0.01	0.67 ± 0.01	0.70 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.01
CT ($\alpha = 0$) (*ours*) *	**0.20 ± 0.01**	**0.38 ± 0.01**	**0.45 ± 0.01**	0.50 ± 0.02	**0.52 ± 0.02**	0.55 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.60 ± 0.02	0.61 ± 0.02
CT (*ours*)	**0.20 ± 0.01**	**0.38 ± 0.01**	**0.45 ± 0.01**	**0.49 ± 0.01**	**0.52 ± 0.02**	**0.53 ± 0.02**	**0.55 ± 0.02**	**0.56 ± 0.02**	**0.58 ± 0.02**	**0.59 ± 0.02**

Lower = better (best in bold)

Experiments: Ablation study

Ablation types

Based on synthetic datasets we evaluate different versions of CT with varying:

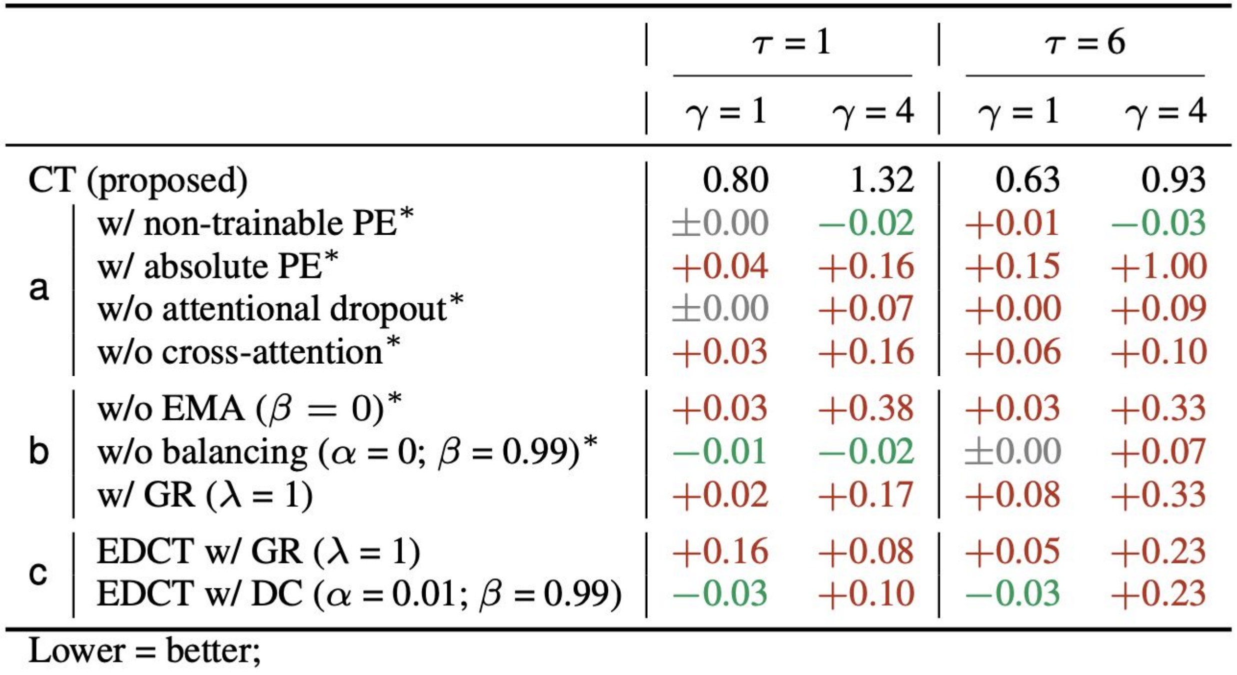
(a) different components within the subnetworks (positional encodings, attentional dropout)

(b) different losses (CDC vs Gradient reversal vs no balancing, w/ vs w/o EMA of weights)

(c) single-subnetwork variant of CT vs original CT

Results

- Combination of **end-to-end three subnetworks architecture and the novel CDC** is crucial (neither work better alone)
- Switching the backbone from LSTM to transformer and using gradient reversal as in (Bica et al., 2020) gives worse results



		$\tau = 1$		$\tau = 6$	
		$\gamma = 1$	$\gamma = 4$	$\gamma = 1$	$\gamma = 4$
CT (proposed)		0.80	1.32	0.63	0.93
a	w/ non-trainable PE*	±0.00	−0.02	+0.01	−0.03
	w/ absolute PE*	+0.04	+0.16	+0.15	+1.00
	w/o attentional dropout*	±0.00	+0.07	+0.00	+0.09
	w/o cross-attention*	+0.03	+0.16	+0.06	+0.10
b	w/o EMA ($\beta = 0$)*	+0.03	+0.38	+0.03	+0.33
	w/o balancing ($\alpha = 0$; $\beta = 0.99$)*	−0.01	−0.02	±0.00	+0.07
	w/ GR ($\lambda = 1$)	+0.02	+0.17	+0.08	+0.33
c	EDCT w/ GR ($\lambda = 1$)	+0.16	+0.08	+0.05	+0.23
	EDCT w/ DC ($\alpha = 0.01$; $\beta = 0.99$)	−0.03	+0.10	−0.03	+0.23

Lower = better;

Conclusion

We proposed a novel, state-of-the-art method: the **Causal Transformer** which is designed to capture complex, long-range patient trajectories.

It combines a **custom subnetwork architecture** to process the input together with a **new counterfactual domain confusion loss** for end-to-end training.

Source Code:
github.com/Valentyn1997/CausalTransformer

ArXiv Paper:
arxiv.org/abs/2204.07258