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, Full Presentation

Institute of AI for Management @ LMU Munich

Who are we?

Valentyn Melnychuk, PhD candidate

Dennis Frauen, PhD candidate

Prof. Dr. Stefan Feuerriegel

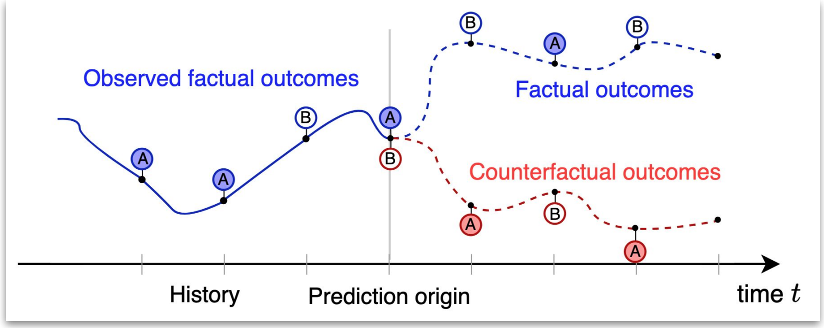
Main research topics: Causal machine learning, Treatment effect estimation, Causal representation learning

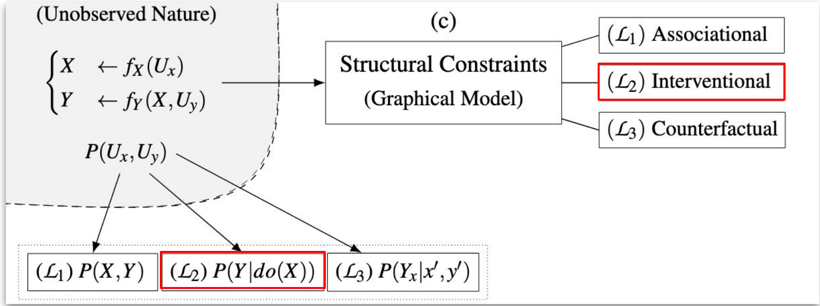
www.ai.bwl.uni-muenchen.de

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 an alternative sequence of treatments, **counterfactual** to a standard treatment policy
- Here, **potential outcomes** are meant, which correspond to the **interventional level of valuation** in Pearl’s Hierarchy of Causal Inference[1]
- 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





[1] Bareinboim, Elias, et al. "On Pearl’s hierarchy and the foundations of causal inference." Probabilistic and causal inference: the works of Judea Pearl. 2022. 507-556.

Introduction: Estimating counterfactual outcomes over time

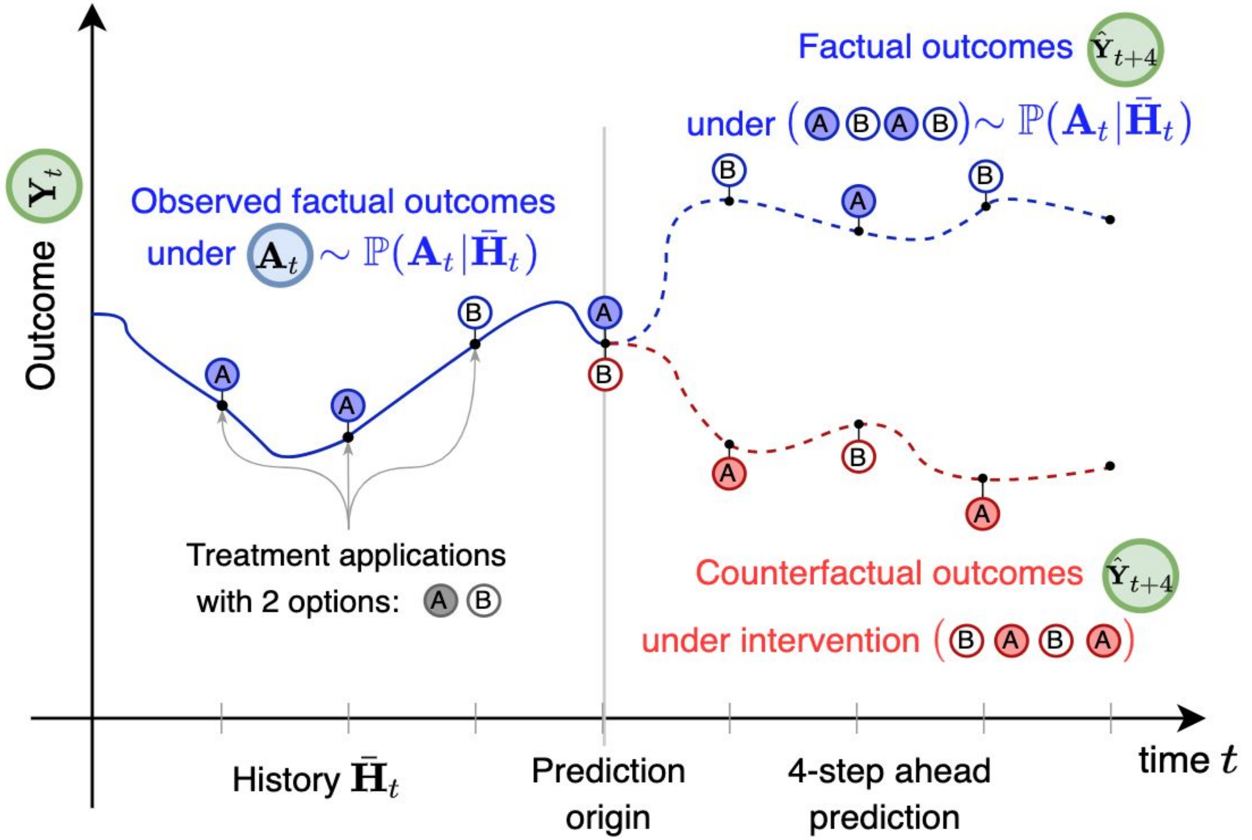
Problem formulation

Given observational dataset of:

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

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

$$\mathbb{E}\Big(\underbrace{\mathbf{Y}_{t+\tau}}_{\tau\text{-step-ahead counterfactual outcome}}\big[\underbrace{\bar{\mathbf{a}}_{t:t+\tau-1}}_{\text{sequence of treatment interventions}}\big] \;\Big|\; \underbrace{\bar{\mathbf{H}}_t}_{\text{history before prediction origin}}\Big)$$



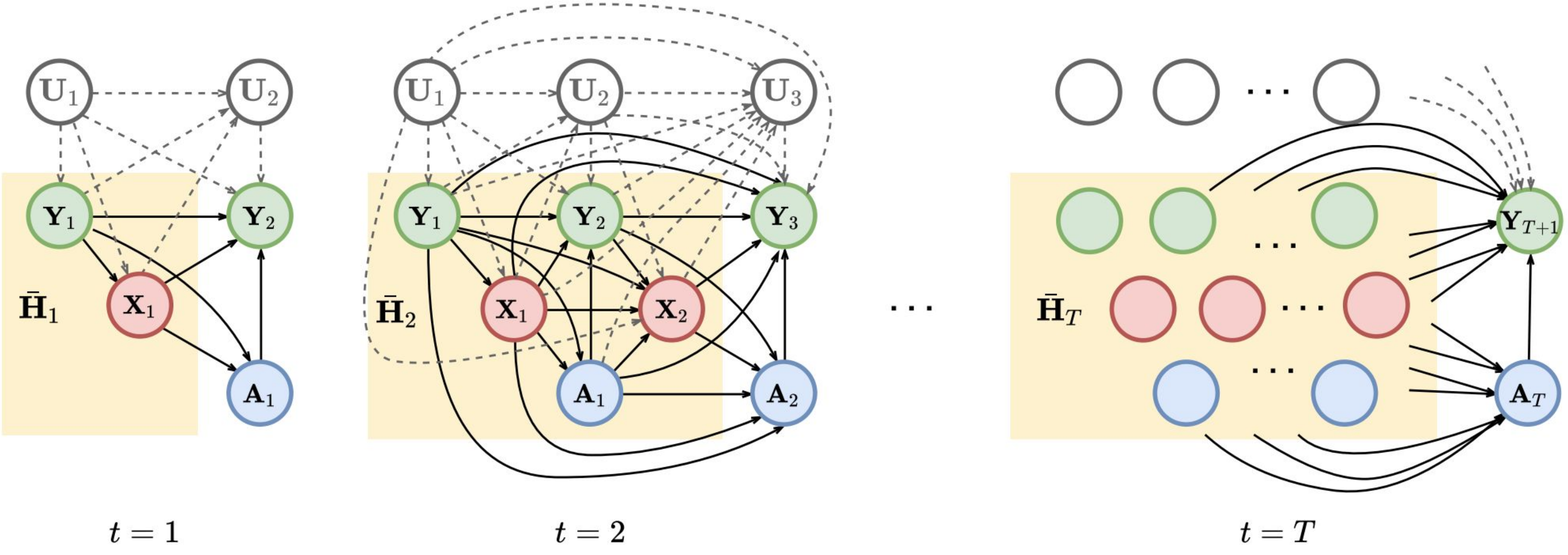
For that, we aim to learn a function $g(\tau, \bar{\mathbf{a}}_{t:t+\tau-1}, \bar{\mathbf{H}}_t)$

*Factual outcomes are observed under standard treatment policy.

Introduction: Assumptions

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 $\mathbf{A}_t \perp\!\!\!\perp \mathbf{Y}_{t+1}[\mathbf{a}_t] \mid \bar{\mathbf{H}}_t$



Introduction: Task complexity

Why estimation is hard?

- Fundamental problem of causal inference:
  counterfactual outcomes are **never directly observed** in a real world
- Traditional machine learning to learn $g(\cdot)$ is either **sub-optimal** (one-step-ahead prediction) or **biased** (multiple-step-ahead prediction) in the presence of time-varying confounding
- Observed **history grows with time**:
  - existing reinforcement literature is non-applicable as this is a **non-Markovian** setting
  - existing literature for cross-sectional setting, e.g. individual treatment effect (ITE) / conditional average treatment effect (CATE), also falls short
- Although the causal effect is identifiable, i.e., with G-Computation formula, it is unclear, how to leverage a **bias-variance tradeoff** and **computational complexity**:

$$\mathbb{E}\left(\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}] \mid \bar{\mathbf{H}}_t\right) = \int_{\mathbb{R}^{d_x} \times \cdots \times \mathbb{R}^{d_x}} \mathbb{E}\left(\mathbf{Y}_{t+\tau} \mid \bar{\mathbf{H}}_t, \bar{\mathbf{x}}_{t+1:t+\tau-1}, \bar{\mathbf{y}}_{t+1:t+\tau-1}, \bar{\mathbf{a}}_{t:t+\tau-1}\right) \times$$

$$\prod_{j=t+1}^{t+\tau-1} \mathbb{P}\left(\mathbf{x}_j \mathbf{y}_j \mid \bar{\mathbf{H}}_t, \bar{\mathbf{x}}_{t+1:j-1}, \bar{\mathbf{y}}_{t+1:j-1}, \bar{\mathbf{a}}_{t:j-1}\right)\, \mathrm{d}\bar{\mathbf{x}}_{t+1:t+\tau-1}\, \mathrm{d}\bar{\mathbf{y}}_{t+1:t+\tau-1}$$

Introduction: Related methods

Related methods

- **Marginal Structural Models (MSMs)** (Robins et al., 2000; Hernan et al., 2001)
  - Base models: linear models wrt. a fixed window taken from history
  - Estimation: (1) propensity score estimation; (2) pseudo-outcome regressions, with IPTW weighted trajectories
- **Recurrent Marginal Structural Networks** (RMSNs) (Lim et al., 2018)
  - Base models: 2 propensity LSTMs, encoder LSTM, decoder LSTM
  - Estimation: (1) propensity score estimation; (2) pseudo-outcome regressions, with IPTW weighted trajectories
- **Counterfactual Recurrent Network** (CRN) (Bica et al., 2020)
  - Base models: encoder LSTM, decoder LSTM
  - Estimation: balanced representations via gradient reversal
- **G-Net** (Li et al., 2021)
  - Base models: time-varying covariates and outcome LSTM
  - Estimation: sampling-based G-computation

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

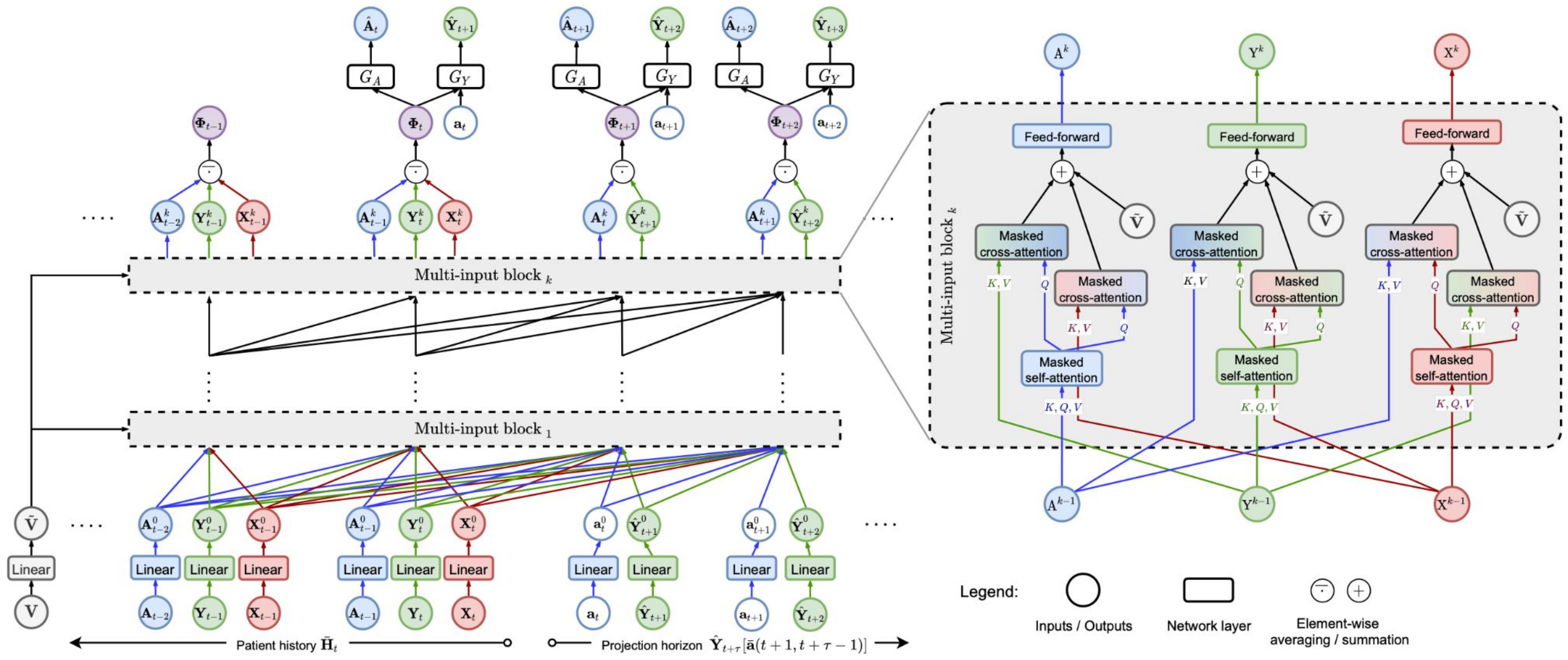
Our contributions

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 adversarial **counterfactual domain confusion (CDC) loss** to address a time-varying confounding
- CT achieves **state-of-the-art performance** on synthetic, semi-synthetic & real benchmarks

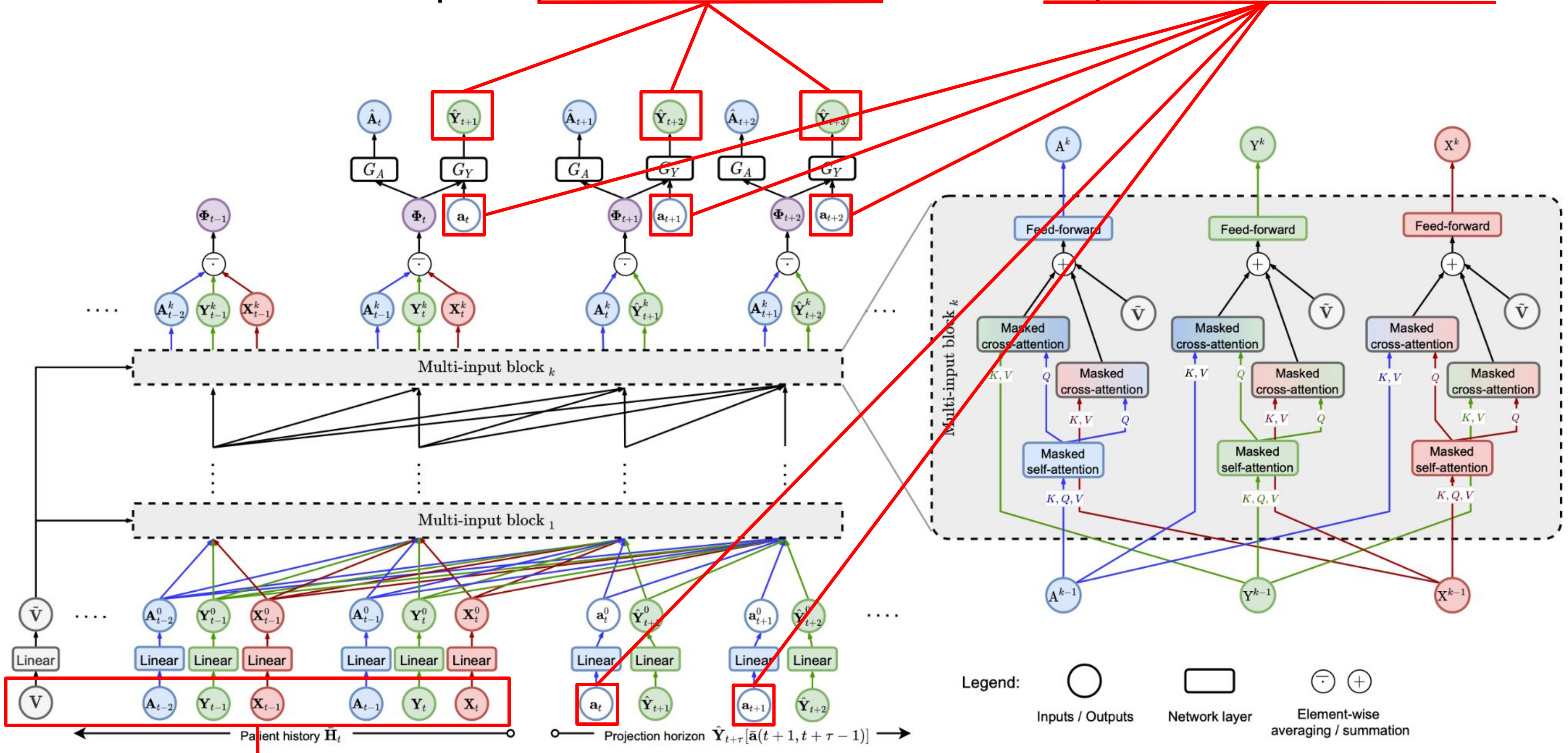
Causal Transformer: Novel architecture

CT is a single end-to-end model for **both one- and multiple-step-ahead prediction**



Causal Transformer: Novel architecture

2. Output – predicted outcomes under a sequence of interventions

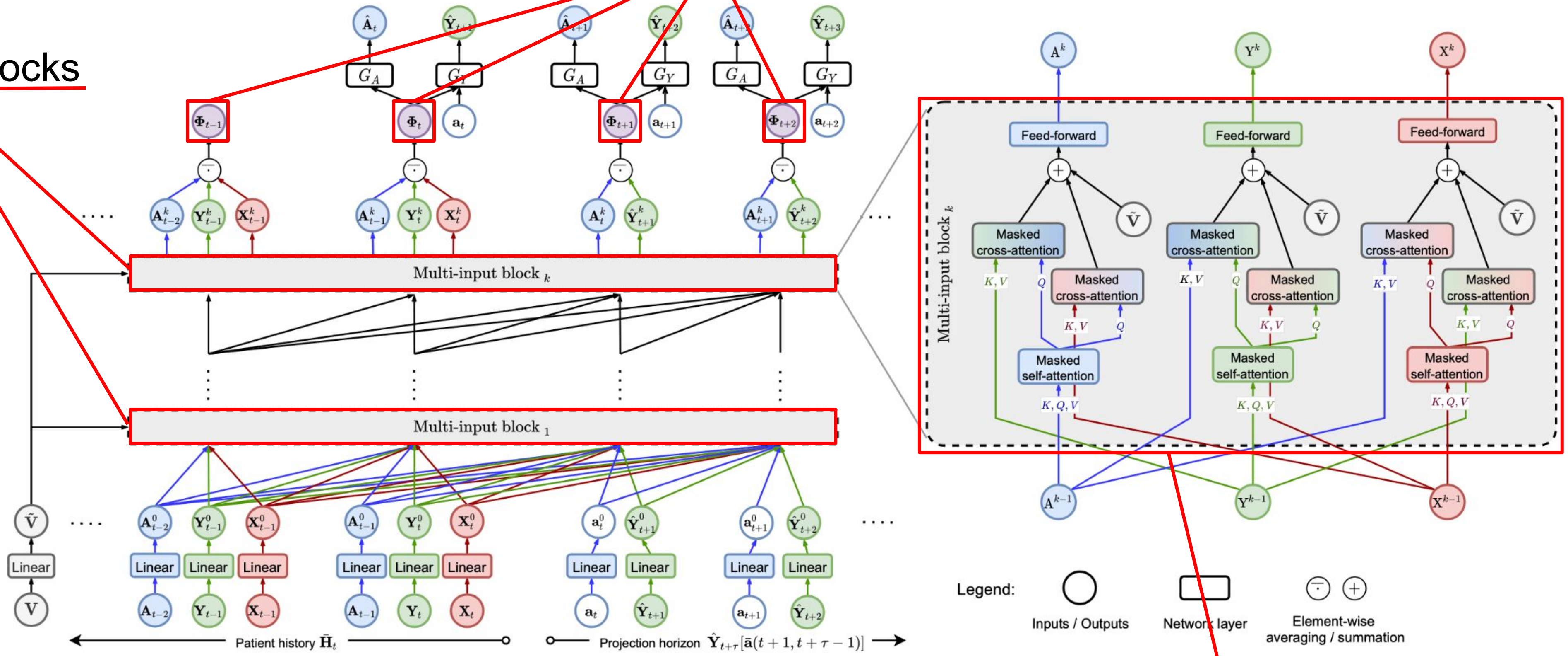


1. Input – observed patient history

Causal Transformer: Novel architecture

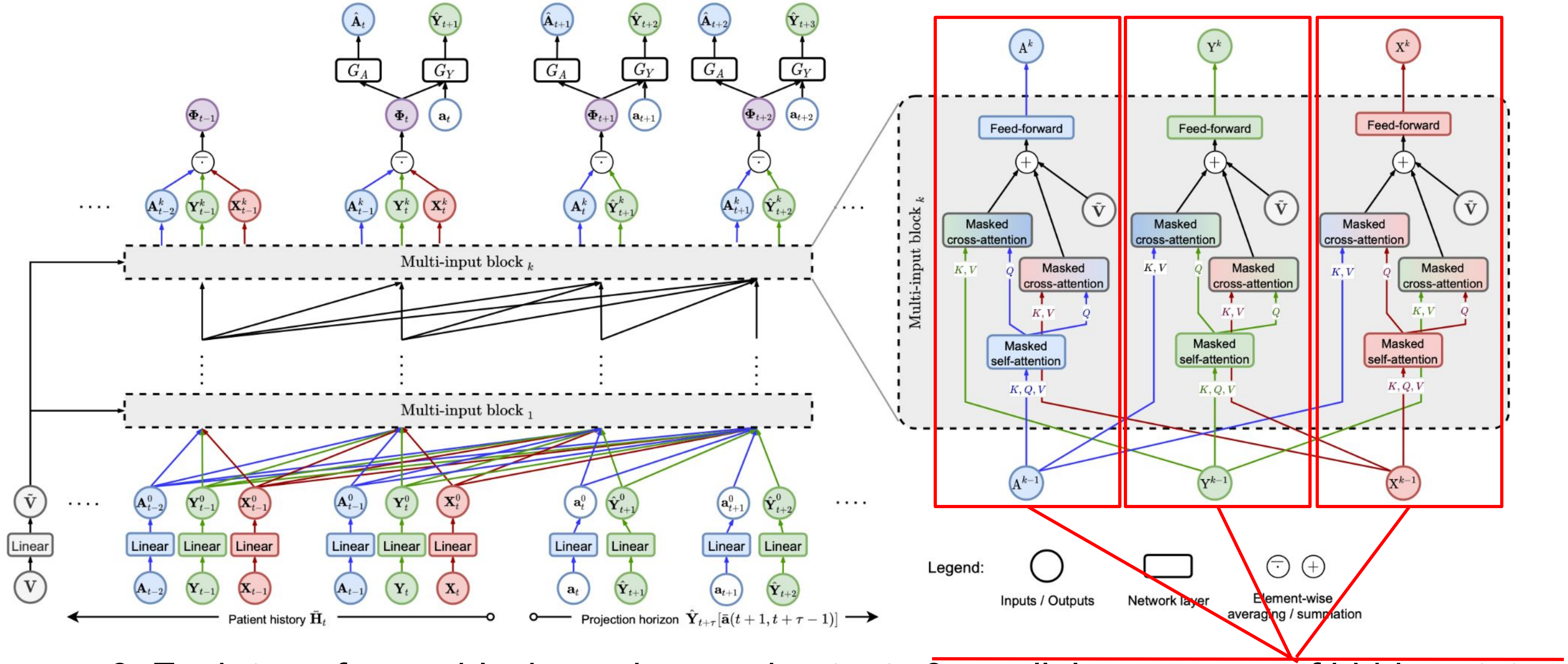
3. Inputs are transformed with a stack of multi-input blocks

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

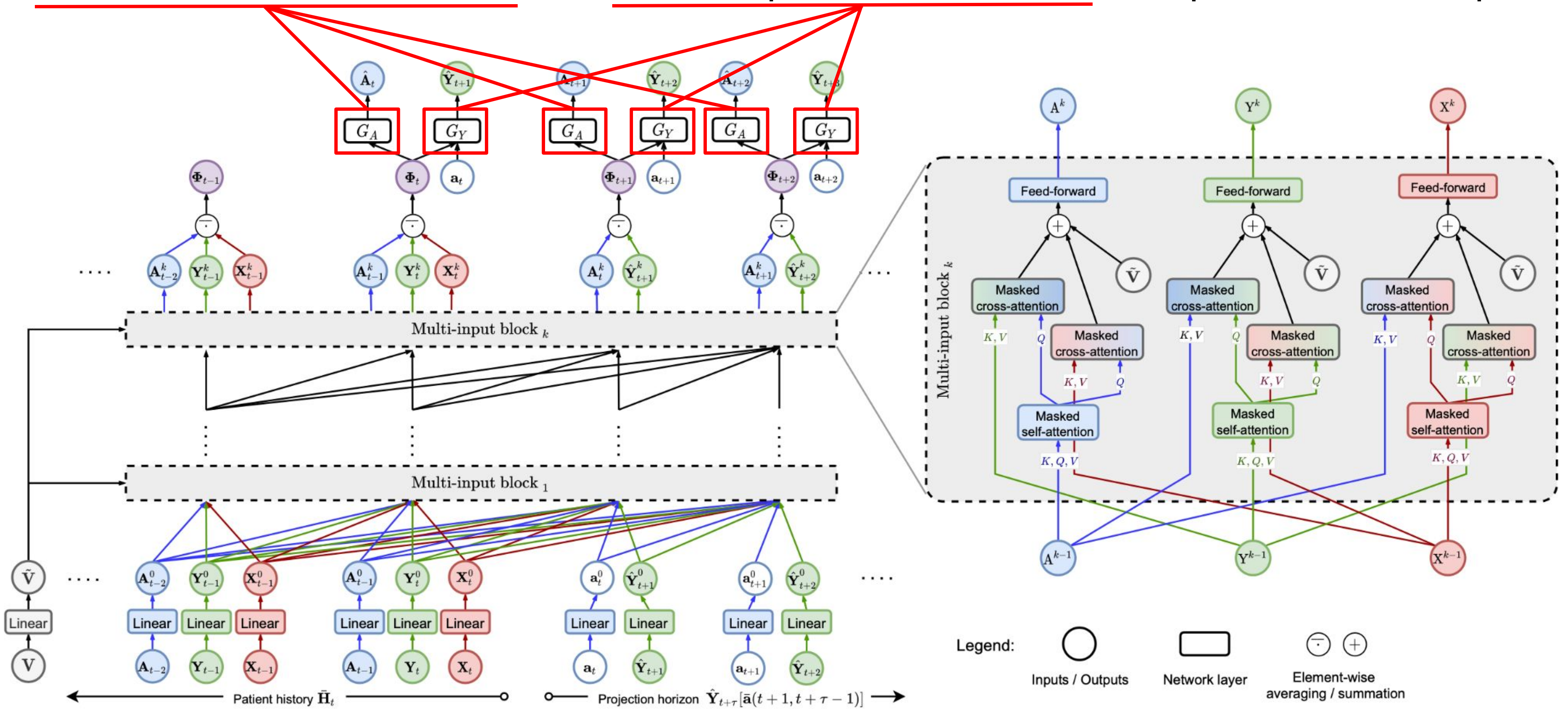
Causal Transformer: Novel architecture



6. Each transformer block receives and outputs 3 parallel sequences of hidden states. I.e., there CT has 3 subnetworks, and the information between them is shared via cross-attentions

Causal Transformer: Novel architecture

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



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

Causal Transformer: Novel architecture

Other details

- Each transformer block is **minimal**[1] and combines
  - (i) multi-head self-/cross-attention with residual connections
  - (ii) feed-forward layer with residual connections
  - (iii) layer normalization
- We employed **attentional dropout**[2], analogously to the recurrent dropout in LSTMs.
- In every self- and cross-attention, we use trainable **relative positional encodings**[3]**,** which:
  - considers the order of treatments, outcomes and time-varying covariates relatively to the prediction origin. E.g., they allow us to distinguish sequences such as, e. g., <treatment A → side effect → treatment B> from <treatment A → treatment B → side-effect>
  - allow for better generalization to unseen sequence length by dropping the order information for the distant past
- **Mini-batch augmentation with masking** is used to enable multi-step-ahead prediction, where future time-varying covariates are unavailable

[1] Dong, Yihe, Jean-Baptiste Cordonnier, and Andreas Loukas. "Attention is not all you need: Pure attention loses rank doubly exponentially with depth." International Conference on Machine Learning. PMLR, 2021.

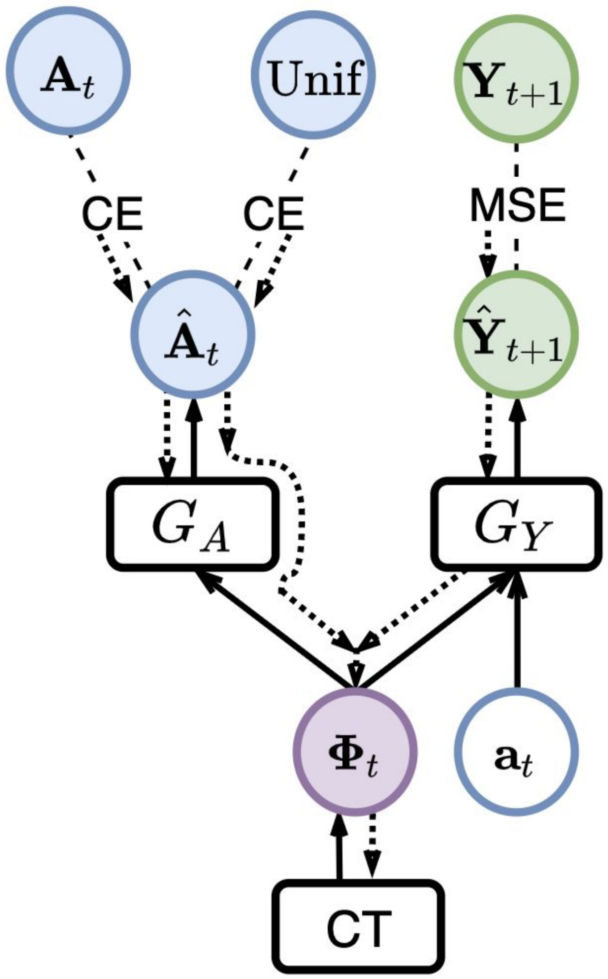
[2] Zehui, Lin, et al. "DropAttention: a regularization method for fully-connected self-attention networks." arXiv preprint arXiv:1907.11065 (2019).

[3] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. "Self-attention with relative position representations." arXiv preprint arXiv:1803.02155 (2018).

Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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



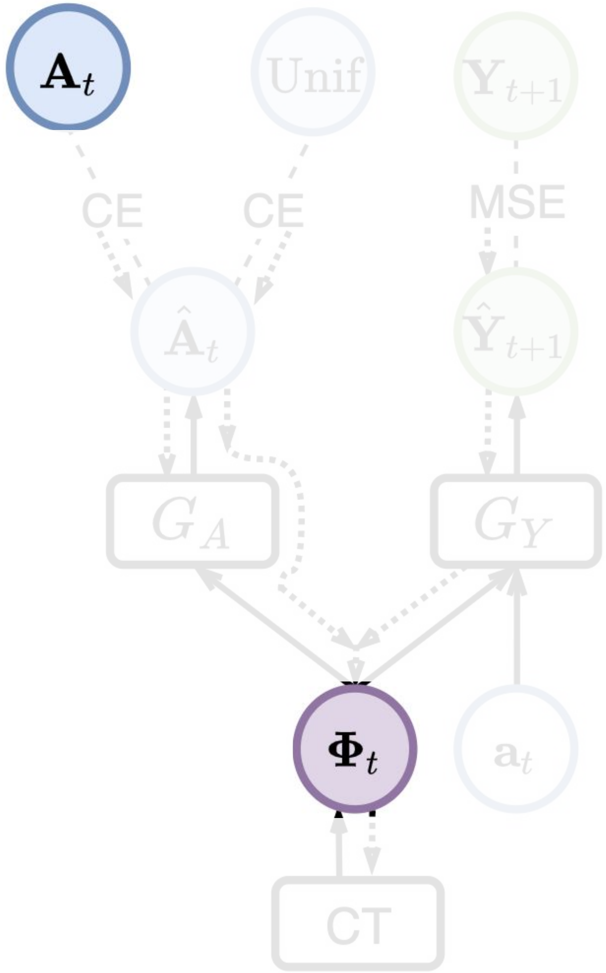
[1] Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." Proceedings of the IEEE international conference on computer vision (2015)

Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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

  (a) make **balanced representations** $\boldsymbol{\Phi}_t$ **non-predictive** of the current treatment $\mathbf{A}_t$



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

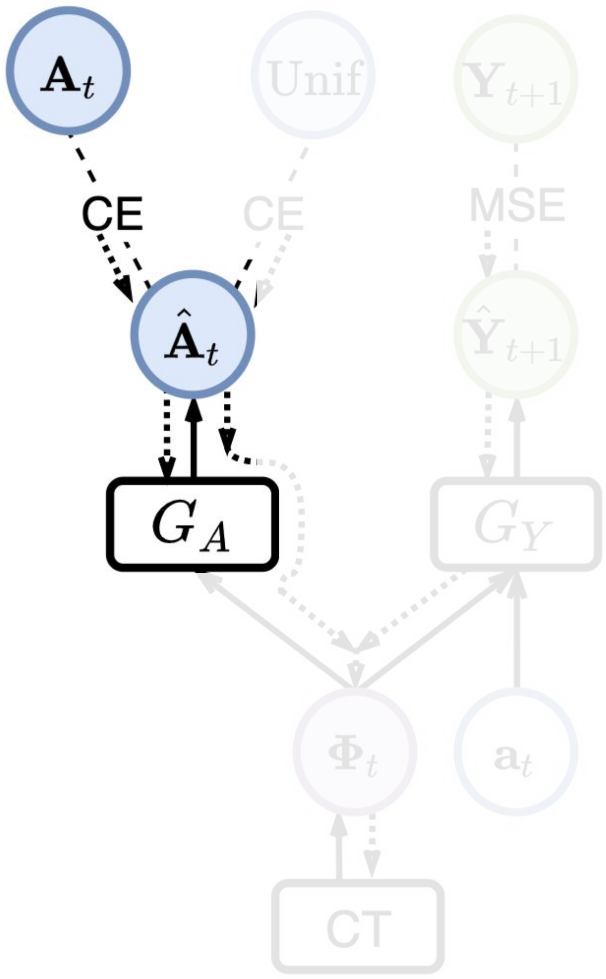
Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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

  (a) make **balanced representations** Φ_t **non-predictive** of the current treatment $\mathbf{A}_t$

  - by minimizing cross-entropy of current treatment wrt. $\boxed{G_A}$



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

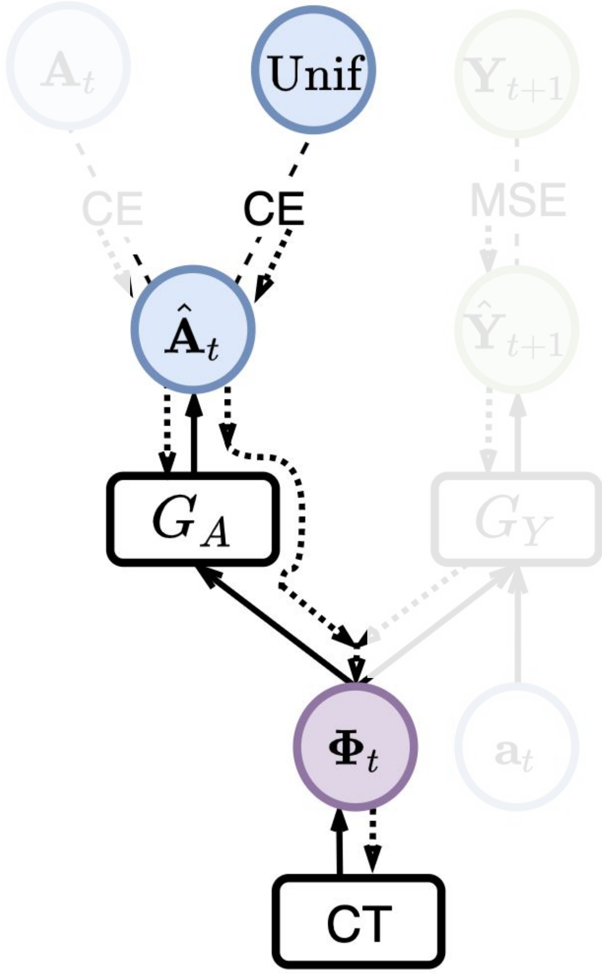
Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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

  (a) make **balanced representations** Φ_t **non-predictive** of the current treatment $\mathbf{A}_t$

  - by minimizing cross-entropy of current treatment wrt. G_A
  - by minimizing cross-entropy between uniform treatment and output of treatment classifier network wrt. CT



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

Causal Transformer: Counterfactual domain confusion (CDC) loss

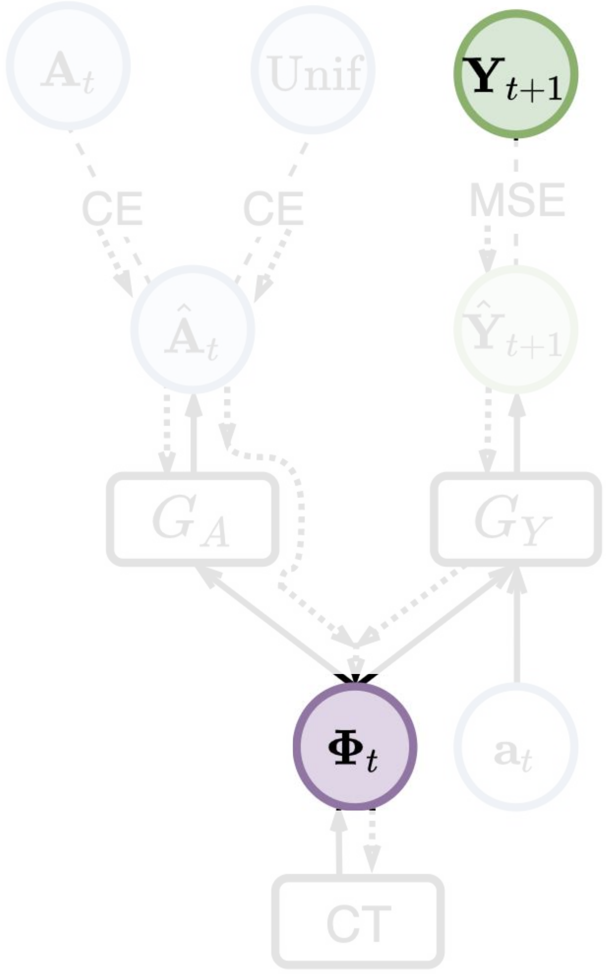
CDC loss

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

  (a) make **balanced representations** $\mathbf{\Phi}_t$ **non-predictive** of the current treatment $\mathbf{A}_t$

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

  (b) make **balanced representations** $\mathbf{\Phi}_t$ **predictive** of the outcome $\mathbf{Y}_{t+1}$



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

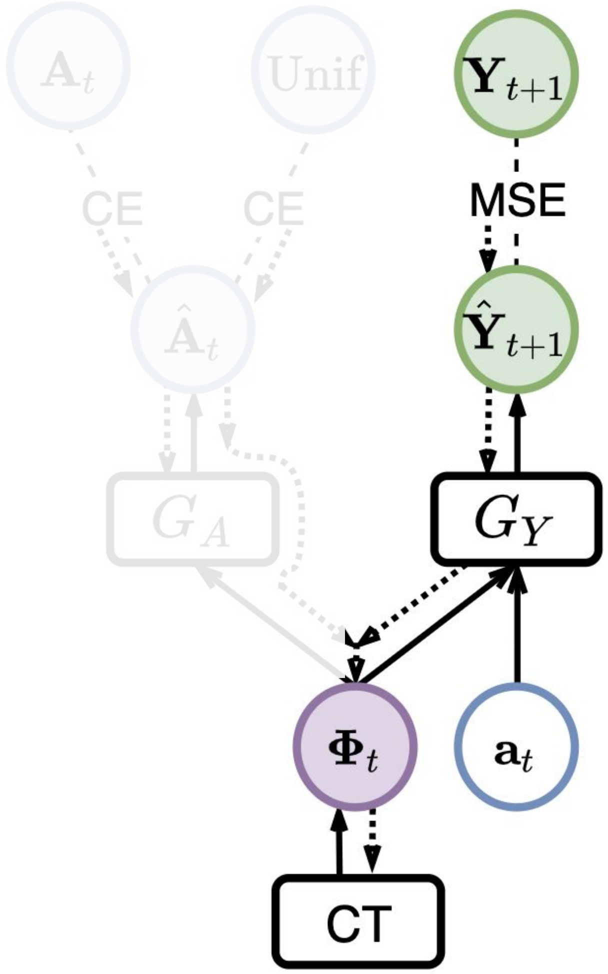
Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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

  (a) make **balanced representations** $\mathbf{\Phi}_t$ **non-predictive** of the current treatment $\mathbf{A}_t$
  - by minimizing cross-entropy of current treatment wrt. $\boxed{G_A}$
  - by minimizing cross-entropy between uniform treatment and output of treatment classifier network wrt. $\boxed{\text{CT}}$

  (b) make **balanced representations** $\mathbf{\Phi}_t$ **predictive** of the outcome $\mathbf{Y}_{t+1}$
  - minimizing factual MSE wrt. $\boxed{\text{CT}}$ and $\boxed{G_Y}$



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

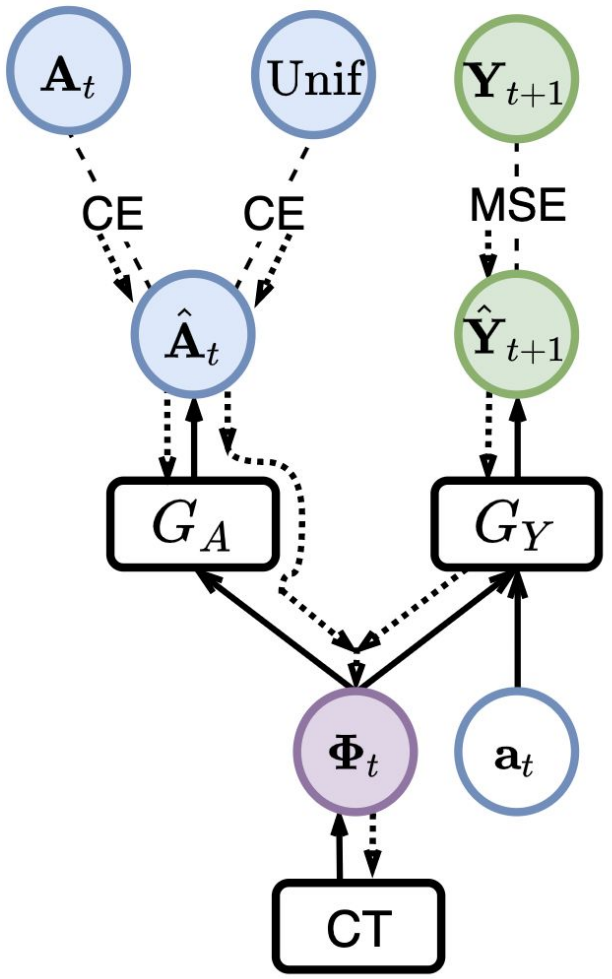
Causal Transformer: Counterfactual domain confusion (CDC) loss

CDC loss

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

  (a) make **balanced representations** Φ_t **non-predictive** of the current treatment $\mathbf{A}_t$
  - by minimizing cross-entropy of current treatment wrt. $\boxed{G_A}$
  - by minimizing cross-entropy between uniform treatment and output of treatment classifier network wrt. $\boxed{\text{CT}}$

  (b) make **balanced representations** Φ_t **predictive** of the outcome $\mathbf{Y}_{t+1}$
  - minimizing factual MSE wrt. $\boxed{\text{CT}}$ and $\boxed{G_Y}$
- Adversarial learning is further stabilized with **exponential moving average** (EMA) of model weights



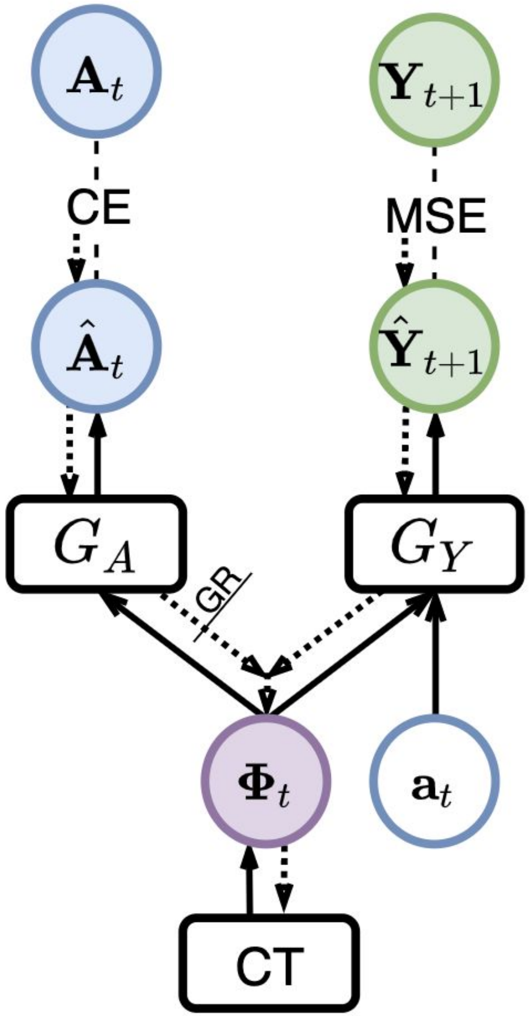
[1] Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." Proceedings of the IEEE international conference on computer vision (2015)

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 vanish, if treatment classifier network learns too fast
- 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

Datasets

- We evaluate CT based on:
  - **synthetic datasets** based on pharmacokinetic-pharmacodynamic model of tumor growth

$$\mathbf{Y}_{t+1} = \left(1 + \rho \log\left(\frac{K}{\mathbf{Y}_t}\right) - \beta_c C_t - (\alpha_r d_t + \beta_r d_t^2) + \varepsilon_t\right)\mathbf{Y}_t \qquad \mathbf{A}_t^c, \mathbf{A}_t^r \sim \text{Bernoulli}\left(\sigma\left(\frac{\gamma}{D_{\max}}(\bar{D}_{15}(\bar{\mathbf{Y}}_{t-1}) - D_{\max}/2)\right)\right)$$

  - self-designed **semi-synthetic dataset** based on MIMIC-III dataset

$$\mathbf{Z}_t^{j,(i)} = \underbrace{\alpha_S^j \, \text{B-spline}(t) + \alpha_g^j \, g^{j,(i)}(t)}_{\text{endogenous}} + \underbrace{\alpha_f^j \, f_Z^j(\mathbf{X}_t^{(i)})}_{\text{exogenous}} + \underbrace{\varepsilon_t}_{\text{noise}} \qquad p_{\mathbf{A}_t^l} = \sigma\left(\gamma_A^l \bar{A}_{T_l}(\bar{\mathbf{Y}}_{t-1}) + \gamma_X^l f_Y^l(\mathbf{X}_t) + b_l\right) \qquad \mathbf{Y}_t^j = \mathbf{Z}_t^j + E^j(t)$$

$$\mathbf{A}_t^l \sim \text{Bernoulli}\left(p_{\mathbf{A}_t^l}\right),$$

  - **real-world dataset** (MIMIC-III)
- 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. For multiple-step-ahead we sampled a fixed number of random counterfactual trajectories

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)

Experiments: Results

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)

- Among all the neural models, our CT has the **smallest runtime**, due to single-stage training procedure with CDC loss and usage of self-attention:

	Main stages of training & inference	Total runtime (in min)
MSMs	2 logistic regressions for IPTW & linear regression	3.5 ± 0.3
RMSNs	2 networks for IPTW & encoder & decoder	109.7 ± 2.3
CRN	encoder & decoder	75.3 ± 17.5
G-Net	single network & MC sampling for inference	118.0 ± 2.0
CT (ours)	single multi-input network	**13.5** ± **4.8**

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 (EDCT) vs original CT

Results

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

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

Extended related work

Method	Setting	Model type (backbone)	Time	Treatments	Framework
HITR (Xu et al., 2016)	DGM (✗)	NP (GP)	Disc & Cont	Seq, Cat	G-computation
CGP (Schulam & Saria, 2017)	C, SO, SI, CSI (✗)	NP (GP)	Cont	Seq, Cat	G-computation
MOGP (Soleimani et al., 2017)	DGM (✗)	SP (GP)	Disc & Cont	Seq, Cont	G-computation
SyncTwin (Qian et al., 2021)	DGM (✗)	SP (GRU-D, LSTM)	Disc	Single-time, Bin	Synthetic control
DCRN (Berrevoets et al., 2021)	C, SO, Cov (✗)	P (3 LSTMs)	Disc	Seq, Bin	Disentangled representation
* MSMs (Robins et al., 2000)	C, SO, SI (✓)	P (Logistic & linear regressions)	Disc	Seq, Cat	IPTW weighted loss
* RMSNs (Lim et al., 2018)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	IPTW weighted loss
* CRN (Bica et al., 2020)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	BR (gradient reversal)
* G-Net (Li et al., 2021)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	G-computation
* *Causal Transformer* (this paper)	C, SO, SI	P (3 transformers)	Disc	Seq, Cat	BR (CDC)

* = Methods with the same assumptions as ours (and thus included in our baselines)

Legend:

- Setting: consistency (C), sequential overlap (SO), sequential ignorability (SI), sequential ignorability but conditional on covariates (Cov), continuous sequential ignorability (CSI), assumed data generating model (DGM)
- Model: parametric (P), semi-parametric (SP), and non-parametric (NP)
- Time: discrete (Disc) or continuous (Cont) time steps
- Treatments: sequential (Seq), binary (Bin), categorical (Cat), continuous (Cont).
- Framework: inverse probability of treatment weights (IPTW), balanced representations (BR)

Attention primer

1. Linear transformations:

$$Q^{(i)} = Q^{(i)}(\mathrm{H}^b) = \mathrm{H}^b\, W_Q^{(i)} + \mathbf{1} b_Q^{(i)\top},$$
$$K^{(i)} = K^{(i)}(\mathrm{H}^b) = \mathrm{H}^b\, W_K^{(i)} + \mathbf{1} b_K^{(i)\top},$$
$$V^{(i)} = V^{(i)}(\mathrm{H}^b) = \mathrm{H}^b\, W_V^{(i)} + \mathbf{1} b_V^{(i)\top},$$

2. Attention weights and scores:

- w/o relative positional encoding $$\mathrm{Attn}^{(i)}(Q^{(i)}, K^{(i)}, V^{(i)}) = \mathrm{softmax}\left(\frac{Q^{(i)} K^{(i)\top}}{\sqrt{d_{qkv}}}\right) V^{(i)}$$
- w/ relative positional encoding $$(\mathrm{Attn}(Q, K, V))_i = \sum_{j=1}^{t} \alpha_{ij}(V_j + a_{ij}^V), \quad \alpha_{ij} = \mathrm{softmax}_j\left(\frac{Q_i^\top (K_j + a_{ij}^K)}{\sqrt{d_{qkv}}}\right)$$

3. Multi-head attention:

$$\mathrm{MHA}(Q, K, V) = \mathrm{Concat}(\mathrm{Attn}^{(1)}, \ldots, \mathrm{Attn}^{(n_h)}).$$

Encoder-Decoder Causal Transformer: Architecture

Two separate transformers, i.e., encoder and decoder, for each task of one- and multiple step ahead predictions