EMPIRICAL STUDY ON THE IMPACT OF NETWORK ARCHITECTURE ON CAUSAL EFFECT ESTIMATION WITH TARNET

AUTHOR: Monika Witczak

AFFILIATIONS:
Supervisor: Rickard Karlsson
Responsible Professor: Jesse Krijthe
Institution: Delft University of Technology

1. INTRODUCTION

- Estimating causal effects from observational data is essential in fields like healthcare, economics, and education.
- Neural networks, such as TARNet [1], can support causal inference by learning representations that reduce confounding.
- While effective, TARNet's architecture (number of layers, neurons) has not been systematically studied.
- This research explores how architectural hyperparameters impact Conditional Average Treatment Effect (CATE) estimation accuracy, aiming to guide model design in causal tasks.

KEY TERMS

- **Causal Inference** - Process of estimating the effect of an intervention or treatment from data where random assignment is not possible.
- **Treatment Effect** - Difference in outcomes between treated and untreated groups.
- **CATE** - Expected treatment effect for individuals with specific characteristics
- **Confounding** - When a variable influences both treatment assignment and the outcome, biasing causal estimates if not properly accounted for.
- **Overlap** (or **Positivity**) - Assumption that every individual has a non-zero probability of receiving each treatment; necessary for valid comparisons between groups.
- **Representation Learning** - Learning transformations of input features to better separate relevant signal (e.g., treatment effects) from noise or confounders.
- **TARNet (Treatment-Agnostic Representation Network)** - A neural network that estimates potential outcomes for treatment and control groups by learning shared representations.
- **TNet** - An architecture that trains separate models for the treated and control groups to estimate potential outcomes.

2. RESEARCH QUESTION

How does varying the hyperparameters, specifically the number of layers and neurons per layer, in a TARNet neural network affect the performance of Conditional Average Treatment Effect (CATE) estimation on simulated datasets?

SUB-QUESTIONS

1. How does TARNet's performance vary across different data regimes (e.g., confounding strength, input dimensionality, and dataset size) when using a fixed architecture?
2. How does the optimal TARNet architecture change in response to these data characteristics?
3. Based on the findings above, what practical recommendations can be made for selecting TARNet architectures under varying data conditions?

3. METHODOLOGY

Literature study → Result reproduction from Shalit et al. and Curth et al. → Experiments with fixed architecture and varying data characteristics, such as sample size and causal structure → Experiments with fixed data settings and varying network architectures (number of layers, neurons per layer) → Analysis and trend identification in different variable and architecture settings and formulation of recommendations

4. FIXED ARCHITECTURE EXPERIMENTS

Fixed architecture: 3 layers with 200 neurons

- **TARNet (SNet1) outperformed the TNet** baseline in almost all settings.
  - **Why?** → Learning a shared representation for control and treatment groups generalizes better than modeling each group separately.
- **TARNet performs better with correlation in the data** (Figure 4), especially when dimensionality increases.
  - **Why?** → Shared layers leverage correlations for compact learning and better generalization.
- **RMSE grew with confounding strength** (Figure 1) or **confounder volume** (= dimensionality, Figure 3)
  - **Why?** → More confounding complicates accurate CATE estimation
- **RMSE formed an inverted U-shape** when increasing the number of confounders within a fixed dimensionality (Figure 2)
  - **Why?** → Initial spike due to introducing confounding, later decline due to reduced noise

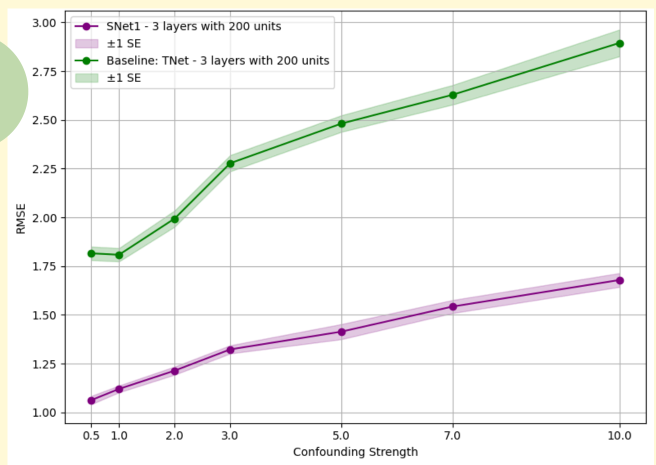

Figure 1: CATE RMSE vs confounding strength for default TARNet architecture with TNet baseline

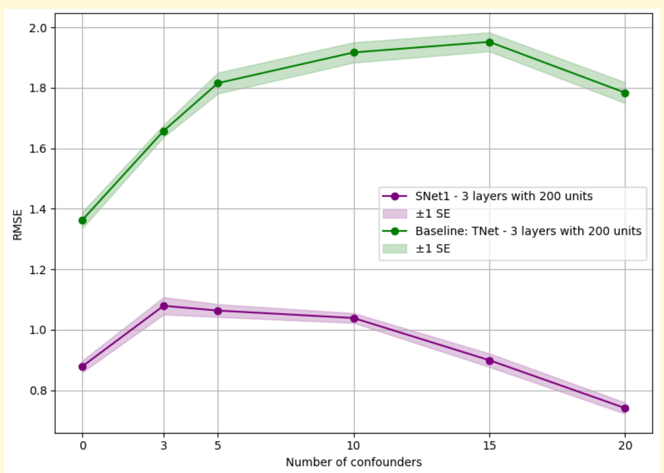

Figure 2: CATE RMSE vs number of confounders in a fixed dimensionality (TARNet and TNet baseline)

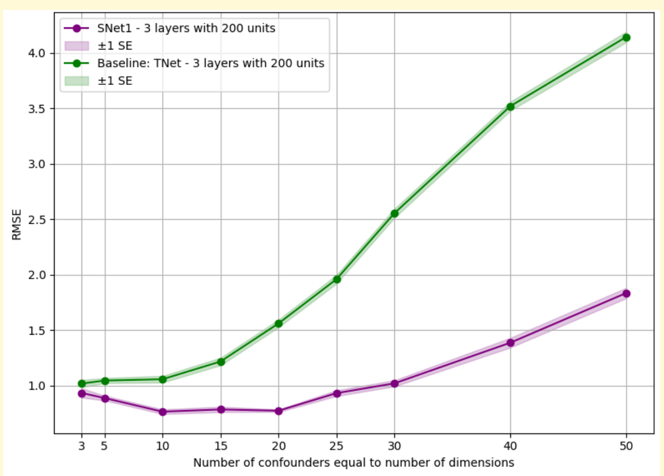

Figure 3: CATE RMSE vs number of confounders equal to dimensionality (TARNet with TNet baseline)

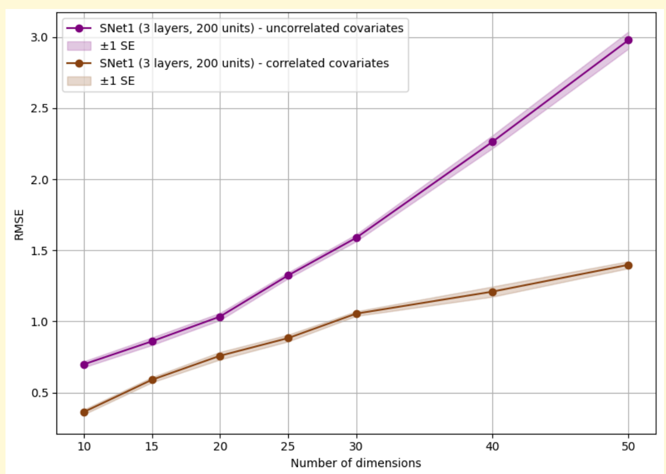

Figure 4: CATE RMSE vs dimensionality for TARNet with correlated and uncorrelated covariates

5. FIXED DATA SETTINGS EXPERIMENTS

A grid of around **30 different architectures** was tested in **fixed data settings**: small sample size, low dimensionality, low signal/confounding, high dimensionality

- **Neurons per layer**: 25, 50, 100, 200, 300, 500
- **Layers**: 1, 2, 3, 4, 5

Table 1: Summary of best-performing TARNet architectures across different data regimes.

Setting	Layers	Neurons	Observation
Small sample size (n = 25, 50, 100)	4 - 5	25	Deep and narrow networks performed best. Shallow models underfit, wide ones overfit.
Moderate sample size (n = 500)	3 - 4	100 - 200	Wider networks became viable, moderately deep and wide architectures performed well.
Low dimensionality (d = 10)	5	100	Deep networks performed well across all variants (confounding only, noise, outcome-relevant).
Treatment heterogeneity in low dimensionality (d = 10)	1 - 2	25, 200	Moderate depth and width offered best generalization; deeper networks began overfitting.
Low confounding strength (ξ = 0.1, 0.3, 0.5)	4 - 5	25 - 100	Deep networks remained effective; narrower widths reduced error in low-signal settings.
High dimensionality, low signal (d = 100)	5	25	Deep and narrow networks resisted overfitting to irrelevant features.
High dimensionality, high signal (d = 50)	5	500	Complex signal best captured by both higher depth and width.

6. FINAL RECOMMENDATIONS

1. Generally use deeper architectures, but consider the context.
2. Adjust network width based on sample size and signal quality.
3. Explore the role of correlated feature structures for TARNet.
4. Consider stability, not just absolute best performance.
5. Avoid unnecessary complexity.
6. Validate empirically on your specific data.

7. CONCLUSIONS

- **No Universal Architecture**: Optimal TARNet architecture is sensitive to the data characteristics, not universal.
- **Deeper is Often Better**: Deeper networks generally outperform shallower ones, especially in high-dimensional or noisy data.
- **Width Depends on Data Quality**: It should adapt to the sample size and noise → narrower for small/noisy data, wider for large/clean data.
- **Stability Trade-Off**: The lowest error doesn't always guarantee model stability; robust, moderately complex models often provide more consistent performance across different neuron counts.
- **Limitations**: Assumes no unobserved confounders; relies on synthetic data (with specific data-generating processes); findings are TARNet-specific; constrained hyperparameter search space.
- **Future Work**: Validate on real-world datasets, compare with other causal models (DragonNet [2], CFRNet [1]), and investigate combined hyperparameter tuning.

REFERENCES

[1] U. Shalit, F. Johansson, D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International conference on machine learning*, 2017, pp. 3076–3085.

[2] C. Shi, D. M. Blei, and V. Veitch, "Adapting Neural Networks for the Estimation of Treatment Effects," Oct. 17, 2019, arXiv: arXiv:1906.02120. doi: 10.48550/arXiv.1906.02120.