HKU Summer Research Proposal

Causal Inference on treatment effect with Deep Learning algorithms

Xu Jiacheng

Nov 15, 2022

1 Summary

This proposal aims at conducting a causal analysis for individual treatment effect by designing a multi-models deep learning algorithm. Inspired by the previous causal inference backbones. We dedicate to incorporating CT images into our covariates dataset and meanwhile expand binary treatment assumption into sophisticated structure-chemical treatment form. To fulfill this idea, we will investigate a novel model, GNNs, to transform chemical structure into a 2D feature map [23], which can be further classified into groups and channeled into subsequent pipelines. We will use the idea of pretrain models, a semi-supervised learning model Simclr and the cross validation method to resolve the problems of lacking sufficient CT images. In the future, we want to test whether transformers [22] can replace the multi-head models we used in causal inference backbones.

2 Background

2.1 Causal inference

Causal inference has long been a fundamental conundrum in Statistics inference. Causal effects, which are defined as the difference of prospective outcomes under various treatments on a common set of units [6], play a key role in analyzing treatment effects from observational data. Missing counterfactual data has been a critical problem for Causal effects analysis. Rubin Causal Model [1] made its contributions by setting up Mathematic equations to define causal effects on medical treatment benefit–there are three major parts: units, treatment and potential outcomes where units represent the information of one certain patient, treatments stand for which treatment patient decides to receive and potential come reflects the patient's situation after certain treatment. To put it simple, Rubin [1] reduced the situation of treatment into binary form and denotes W to imply weather a patient receive the treatment: W=0 indicates the control group while W=1 indicates the treatment group ; Y(0) is the outcome if the patient belongs to control group and Y(1) is similar defined. The individual causal effect,which we use for measuring the treatment benefits on certain individuals, is then defined as the difference between two possible outcomes.

The main problem of Causal effects analysis is that we cannot have the both data of Y(0) and Y(1) at the same time and our estimation on the missing data part can be biased, in statistical terminology, the missing counterfactual data problem.In history, statisticians employ the traditional statistical method to eliminate the bias between control group data and treatment group data. Propensity score matching has been a fundamental traditional statistical method to achieve balanced data groups and solving the missing counterfactual data problem on an initial basis [17]. However, recent scholar [9] points out propensity score matching cannot perform well in high dimensional covariate data and has its limitation in dealing with high dimensional clinical tabular data. With the development of machine learning and AI technology, famous statisticians like Judea Pearl [13] suggested that machine learning can be applied to deal with balancing data, make up for the missing counterfactual data and train treatment assignment models. That idea guided me to design advanced deep learning algorithms to resolve causal inference problems in the future.

2.2 Deep learning algorithms

Deep learning, the multiple-layer model that learns abstract representations and features from various types of data, has been popular since AlexNet [10] demonstrated its unparalleled power in image classification. Deep learning models have a wide range of applications in image classification, object detection, speech recognition, segmentation, etc. Recent years have witnessed a bunch of innovative and revolutionary landmark models designed for solving issues in various domains: Resnet [5] and VGG [19] are prominent models for image classifications and improve the overall accuracy to the next level. Transformers [22] play a key role in speech recognition by innovating self-attention and mutual attention mechanisms. GNNs [18] even provide a robust solution for data in the graph nature, incorporating a data form into deep learning based analysis. These landmark models inspire my research and lay a solid foundation for my future research.

Deep learning methods' application in the healthcare area has been an overwhelming trend in the past decade. Convolutional network based algorithms have been developed to analyze treatment effects. In Jiang's work [7], they employed a Convolutional network resembling Resnet [5] to learn a survival score from CT images to predict treatment benefits. Cui's research [2] also designed Convolutional networks to abstract features from MRI and combine it with statistical models like the LASSO logistic regression model. GNNs [18] make the contribution through employing a chemical language and information system, SMILES [20], to transform chemical structures and drugs as graphs and then abstracting useful representative features from graph based data. Specifically, GNNs based models are introduced to conduct analysis on the critical relationship between human microbes and drug associations [11]. Furthermore, by employing the concept of contrast learning, GNNs based algorithms successfully embedded vast types of drug into a representative feature map, transforming structured data into 2D image data and visualizing the embedding space [23]. With these applications, 2D image data and structure data can be transformed into 1D statistical representative features in the first step and then be channeled into subsequent causal inference pipeline,combined with statistical survival prediction models like cox proportional hazards model, to conduct a robust causal effect prediction on treatments. The development of deep learning algorithms lays a solid theoretical foundation and builds up a tremendous platform for my future PhD study.

3 Literature Review

Causal inference and deep learning models have developed at an amazing speed over the past 2 decades. With Rubin [1] defining the preliminary for causal inference model and propensity score matching method proves to be limited to low dimensional situations [9], Judea Pearl points out the significance of applying deep learning algorithms to the causal inference area [13]. The vast landmark models including Convolutional networks [5,10,19], Transformers[22] and GNNs [18] allow us to abstract representative features from CT images and structured chemical drugs and channel them into subsequent causal analysis pipelines.

The literature on the interdisciplinary application of deep learning algorithms on causal analysis of treatment effects is vast. Farajtabar [3] added a special MMD (Maximum Mean Discrepancy) loss to balance the distribution of two groups of data and then designed a multi-head algorithm to train data from controlled groups and treatment groups separately. Yao's team [25] tried to learn a balanced representative space for original imbalance data li by adding two special concept related loss, position-dependent deep metric (PDDM) and middle point distance minimization (MPDM), to the deep learning algorithms. Both scholars conducted experiments with Infant Health and Development Program (IHDP) datasets, a randomized experiment dataset which contain counterfactual data. However, they are both limited to binary treatment setting and linear pretreatment covariates, and cannot handle structured drugs treatment and CT image covariates.

Realizing the limitations existing in the previous two essays, there is a growing number of literature focusing on learning representative features from CT images to conduct causal analysis on treatment effects. Both Jiang's [7] and Cui's [2] work was dedicated to learning a survival rate score from CT/MR images, combined with several linear clinical characteristics of patients to conduct a basic treatment effect prediction. While Cui's [2] team only employed statistical models like the LASSO logistic regression model, Jiang [7] conducted a basic causal inference analysis with

the traditional propensity score matching method and further employed cox proportional hazards model and Kaplan-Meier lines to test and supervise the acquired scores. They both used CT images and clinical data from prestigious hospitals and had obvious limitations in lack of deep learning based causal inference models, which are the prerequisite of statistical robustness.

Aware of the defect in previous treatment causal effects prediction that they were not able to conduct predictions on multiple types treatment and structured drugs treatment, researchers these years concentrated on expanding the binary treatment input form to a continuous form or a graph form. Kaddour [8] successfully involved the structured drug treatment into causal effect analysis. They referred to Robinson's Decomposition method [16] to split the whole prediction model into a product of two separate models, predicting treatment effects and pretreatment covariates respectively. For the structured treatment prediction part, they both did experiments on GNNs and other baseline models to compare its accuracy. They use real life drugs from the QM9 [14] dataset as treatment input data and gene expression from a real dataset TCGA [21] as covariates input data, transforming chemical structure into graphs by referring to a chemical language and information system, SMILES [20], and then learning representative features from each graph. Based on Kaddour's study, Zhang [26] improved the model by replacing the original Multilayer perceptrons model in the causal inference backbone with Transformers [22]. What is innovative is that they take the treatment dosage into consideration and add an extra dimension into the treatment input. They also simulate a synthetic dataset including 1000 Watts–Strogatz small world graphs [24] to test the performance of their models.

Thought methods from Kaddour [8] and Zhang [26] are relatively advanced than previous models in involving more complex treatment situations, there is still improvement space for future study in this area: they still only included linear clinical characteristic into pretreatment covariates and did not use CT images as input; They still failed to exclude common confounding factors for covariates and treatments. In future studies, it is hoped that we can both include CT images as covariates and involve complex structured treatment form together to construct a more robust causal effects prediction on treatment. Meanwhile, we can refer to previous work which introduced instrumental variables [4] or proxy variables [12] to exclude the negative effects of common confounding factors for covariates and treatments.

4 Objective

- Design a deep learning pipeline which can transform CT image and structured chemical therapy into 1D representative feature and then combine acquired features with clinical tabular data to channel into a subsequent causal inference pipeline
- Future design a subsequent causal inference pipeline to eliminate the bias between different treatment groups and balance the distribution of the whole dataset so as to accurately estimate the causal effect of certain treatment
- Improve the efficiency of the model and conduct analysis with limited data. Tackle the problems that result from lack of CT image data
- Design or employ proper statistical survival rate prediction model and evaluation metrics to test and supervise the process of transforming CT images and structured chemical drugs.

5 Methodology

On the first step, We mainly plan to refer to Kaddour's [8] method to transfer structured treatment into linear data and Jiang's [7] method to learn representative features from CT images. Then we combine clinical pretreatment covariates with acquired information from CT images and structured treatment to conduct a causal effect prediction with a deep learning backbone referring to Farajtabar's [3] work.

5.1 Preliminary

We refer to the preliminary in Kaddour [8] to define our problem setting. The observed data D consists of n triplets $\{x_i, t_i, y_i)\}$, $i = \{1, 2, ..., N\}$, where x_i is the feature or pretreatment covariates of the i_{th} unit, t_i is the treatment (intervention), and y_i is the potential outcome.To put it simple, we focus on m possible integer values for treatment where $t_i = \{0, 1, 2, ..., M\}$. The potential outcome for patient i if the patient chooses the treatment with assigned value m is Y_i^m. The observed outcome can thus be expressed as:

$$yi = \sum_{m=0}^{M} Y_i^m * m \tag{1}$$

For the i-th patient's pretreatment covariates xi and assigned treatment value m, we are interested in estimating the individual impact of the treatment $im(x_i, m)$, the difference between the potential outcome of m-th treatment group and that of controlled group (or individual treatment effect (ITE)):

$$im(x_i, m) = Y_i^m - Y_i^0 \tag{2}$$

5.2 Deep learning pipelines

The whole pipeline is divided into three major parts: CT image feature extraction, Structured treatment embedding and Causal Inference backbones. We respectively gain design inspiration from several previous papers [3, 7, 8]. As for code work, I will employ Matplotlib and Wandb, two Python plotting libraries, for demonstration and visualization results. Also, we follow a Pytorch Lightning format to organize our code.

5.2.1 CT image feature extraction

We mainly follow the idea of Jiang [7] to extract our image features and we design our learned features to a dimensional score which is in accordance with survival possibility of each patient. We then focus on testing the optimal threshold to divide the whole cohorts into high, median and low risk groups so that we can better supervise the accuracy of feature extraction. We basically use Resnet [5] and other landmark models as backbones for this part and focus on designing novel loss functions to replace original ones in these models. An example of CT image feature extraction is shown in Figure 1.

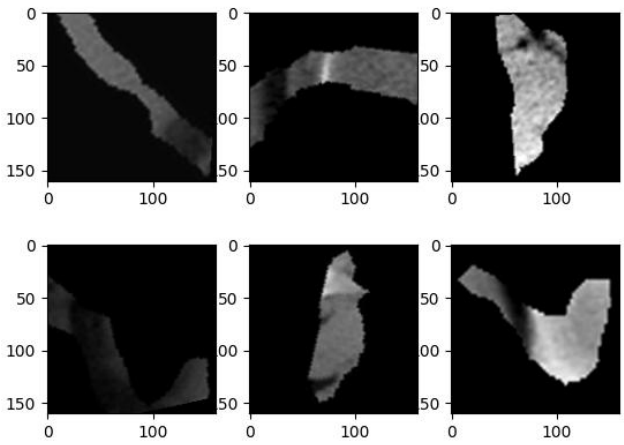

Figure 1: A set of CT images after being cropped by masks

5.2.2 Structured treatment embedding

We mainly refer to Kaddour [8] and concentrate on transforming structured treatment into continuous linear treatment value in the first stage and then employ a fully connected neural network to assign and classify these continuous treatment values into M groups. With the whole dataset grouped into M treatment groups and one controlled group, we are able to train a multi-head causal inference model for each group respectively. The sample 2D feature map is provided in Figure 2.

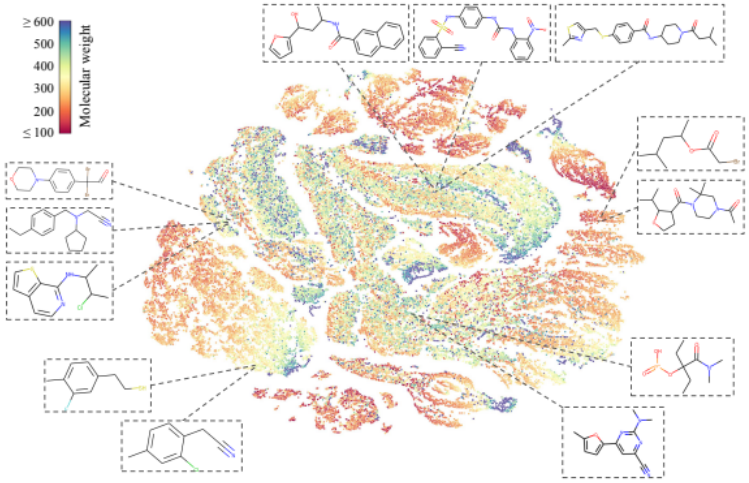

Figure 2: A 2D feature map for structured chemical

5.2.3 Causal inference backbones

In Farajtabar's [3] work, they design a causal inference backbone with two heads to balance the distribution of the dataset. In our work, since we expand our treatment space from binary to M+1 possible values(M treatment groups +one controlled group), we design our model with M+1 heads and train these heads respectively. As for the input data, we combine the learned features from CT images and pretreatment covariates. For Hyperparameter M, we basically determine its value by the results from validation sets. In the future, we also would like to test whether transformers have a better performance than the multi-head models in the future research. A special case when M=2 is shown in Figure 3.

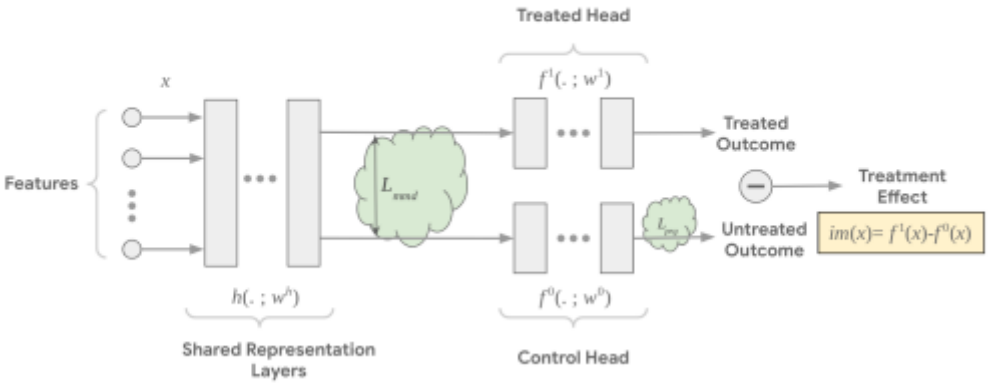

Figure 3: A two-head model for causal inference backbones

5.3 Datasets and Experiments

Our data were collected from 540 patients who received complete or partial radical gastrectomy at Nanfang Hospital at Southern Medical University in Guangzhou, China, between January 2007 and December 2014. Each patient has a 2D or 3D CT image and a bunch of traditional clinical tabular data.

5.3.1 Solutions for a limited data volume

Dealing with the problem of a shortage of data, as CT image data is expensive and scarce, is the most challenging part. The model that we initially intended to train is overfit and did not achieve high prediction accuracy. The current solution I have is through two major kinds of solutions. One is to base on landmark models such as Resnet [5], loading pretrained model parameters, freezing part of layers. Another is to base on semi-supervised learning models such as Simclr. In addition, we also employ the cross validation method to save data which used to be reserved for test sets. By overcoming the constraints of input data, I was able to take the model's efficiency to the next level and make greater use of limited data.

5.3.2 Loss function and metrics

In the CT images feature extraction pipeline, we design our loss function to be the cox loss function [15], which is statistical model based loss specializing in survival prediction. In the Causal Inference pipeline, we mainly use a loss function based on the concept of MMD(maximum mean discrepancy), which is innovated to measure the discrepancy between two distribution(groups of data) by projecting the data into a certain dimensional space called Reproducing Kernel Hilbert Spaces and adopting a kernel method.

5.4 Statistical models

We use statistical models to assist the deep learning models and supervise our training results with a more direct and tangible method. Continuing with our work in training a survival prediction score from the CT images and separating them into three risk level groups, we plot Kaplan-Meier lines for each group to see whether our learned score effectively distinguishes people's risk level. A km lines sample is shown in Figure4.

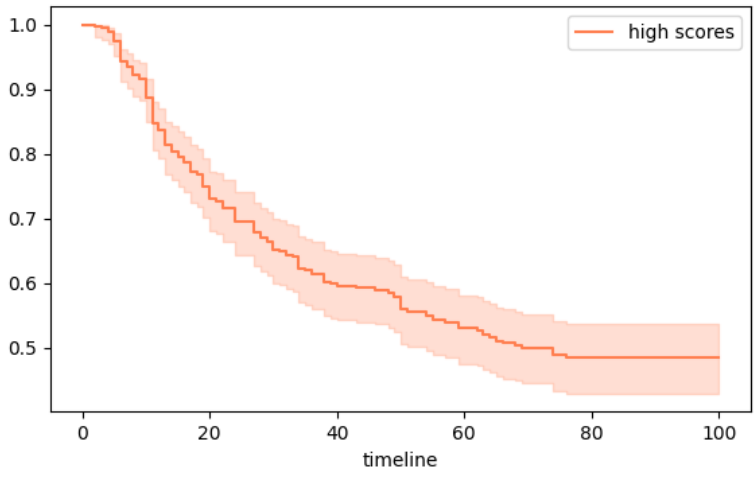

Figure 4: Kaplan-Meier lines for high-score group

6 Significance

- **Up to ethical standards and respect human rights**: To eliminate bias resulting from observed datasets, scientists used to conduct RCTs(Randomized controlled trials) to balance the distribution between the treatment group and the controlled group. However, apart from the expensive cost, RCTs are likely to encounter ethical problems and challenge human rights [25]. Causal Inference models provide an alternative unbiased solution for observation study.
- **Save experts' time and improve efficiency**: Traditional cancer treatment prognosis and diagnosis tend to require an expert team to spend a huge amount of time to provide an authoritative and trustworthy version. With the assistance of deep learning based tools, experts no longer need to manually design masks for CT images and can provide service for more patients efficiently.
- **Provide trustworthy and economical diagnosis for average people**: Since some medical problems like Cancer treatment are difficult to be fully interpreted by medical and biochemical knowledge. It is risky to solely rely on the conclusion from medical and biochemical aspects without considering statistical robustness. Causal analysis provides people with more trustworthy and economical choices for treatments.
- **Facilitate and take advantage of the Hong Kong medical industry**: Hong Kong has one of the best medical systems in the world and biomedical related quantitative research is quite mature and promising. The data source is more reliable and adequate, which makes it more likely to bring theoretical results into clinical applications and facilitate the development of digital healthcare in Hong Kong

References

[1] Causal Inference Using Potential Outcomes: Design, Modeling, Decisions: Journal of the American Statistical Association: Vol 100, No 469. (n.d.). Retrieved November 22, 2022, from

[2] Cui, Y., Yang, X., Shi, Z., Yang, Z., Du, X., Zhao, Z., & Cheng, X. (2019). Radiomics analysis of multiparametric MRI for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. European Radiology, 29(3), 1211–1220.

[3] Farajtabar, M., Lee, A., Feng, Y., Gupta, V., Dolan, P., Chandran, H., & Szummer, M. (2020). Balance Regularized Neural Network Models for Causal Effect Estimation (arXiv:2011.11199). arXiv.

[4] Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. Proceedings of the 34th International Conference on Machine Learning, 1414–1423.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 770–778.

[6] Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), 945–960.

[7] Jiang, Y., Jin, C., Yu, H., Wu, J., Chen, C., Yuan, Q., Huang, W., Hu, Y., Xu, Y., Zhou, Z., Fisher, G. A., Li, G., & Li, R. (2021). Development and Validation of a Deep Learning CT Signature to Predict Survival and Chemotherapy Benefit in Gastric Cancer: A Multicenter, Retrospective Study. Annals of Surgery, 274(6), e1153–e1161.

[8] Kaddour, J., Zhu, Y., Liu, Q., Kusner, M. J., & Silva, R. (2021). Causal Effect Inference for Structured Treatments. Advances in Neural Information Processing Systems, 34, 24841–24854.

[9] King, G., Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27(4), 435–454.

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

[11] Long, Y., Wu, M., Kwoh, C. K., Luo, J., & Li, X. (2020). Predicting human microbe–drug associations via graph convolutional network with conditional

[12] Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., & Muandet, K. (2021). Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction. Proceedings of the 38th International Conference on Machine Learning, 7512–7523.

[13] Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54–60.

[14] Quantum chemistry structures and properties of 134 kilo molecules — Scientific Data. (n.d.). Retrieved November 22, 2022, from

[15] Regression Models and Life-Tables—Cox—1972—Journal of the Royal Statistical Society: Series B (Methodological)—Wiley Online Library. (n.d.). Retrieved November 22, 2022, from

[16] Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. Econometrica, 56(4), 931–954.

[17] ROSENBAUM, P. R., & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

[18] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. IEEE Transactions on Neural Networks, 20(1), 61–80.

[19] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition (arXiv:1409.1556). arXiv.

[20] SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules — Journal of Chemical Information and Modeling. (n.d.). Retrieved November 22, 2022, from

[21] The Cancer Genome Atlas Pan-Cancer analysis project — Nature Genetics. (n.d.). Retrieved November 22, 2022, from

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30.

[23] Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), Article 3.

[24] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), Article 6684.

[25] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018). Representation Learning for Treatment Effect Estimation from Observational Data. Advances in Neural Information Processing Systems, 31.

[26] Zhang, Y. F., Zhang, H., Lipton, Z. C., Li, L. E., & Xing, E. P. (2022). Can Transformers be Strong Treatment Effect Estimators?. arXiv preprint arXiv:2202.01336.