Deep Mendelian Randomization: Using Mendelian Randomization to Detect Learned Causal Relationships in Deep Learning Models

Anonymous Authors[1]

Abstract

1. Introduction

Recently, deep learning models have been used to classify genomic features such as transcription factor binding, chromatin accessibility, the presence / absence of histone marks, and RNA binding protein binding . These models achieve high predictive accuracy on these tasks and learn feature detectors that match experimentally verified features . Furthermore, multi-task models such as DeepSEA achieve high accuracy simultaneously on multiple genomic feature prediction tasks. One question we can ask about these multi-task models is whether, through learning to predict multiple features jointly, they learn experimentally determined causal relationships between these features.

To try and answer this question, we apply Mendelian randomization, an instrumental variable approach for causal inference, to the problem of detecting learned causal effects in genomic deep learning models. Our algorithm obtains local (sequence level) and global (genome level) estimates of the linear causal relationship between two biological processes learned by a multi-task genomic prediction model. In this work, we apply our approach to estimating the learned causal effect of transcription factor binding on chromatin accessibility in a single cell type, but our method can in principle be applied to other processes that are believed to satisfy the instrumental variable assumptions.

2. Background & Related Work

2.1. Deep Learning Interpretability

2.2. Deep Learning Model Uncertainty

2.3. Mendelian Randomization

3. Methods

3.1. Algorithm Overview

Our algorithm attempts to determine whether data a trained model generates reflects known causal relationships in the underlying data-generating process. It requires as input a trained model[1] and a set of one-hot encoded sequences for the model to make predictions on.

Given this, it outputs a set of local, sequence-specific and exposure-specific causal effects and set of global, exposure-specific causal effects. It accomplishes this (see for visual depiction) via the following steps for each exposure:

1. Randomly samples sequences to make predictions for the exposure and outcome on (“reference sequences”).
2. Perform *saturation in-silico mutagenesis* for each reference sequence to generate sequence length $\times$ number of nucleotides -1 mutated sequences per original sequence.
3. For each reference and set of mutated sequences, use MC-dropout to generate predictive means and standard errors of binding probabilities for the (reference—mutated) sequences.
4. Generate sequence length $\times$ number of nucleotides -1 *effect sizes* by taking the differences between each mutated sequence’s predictive mean and the corresponding reference sequence’s predictive mean and the standard error of these differences.
5. Apply Mendelian randomization to each reference sequence’s effect sizes and their standard errors to es-

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

[1] The model could in principle be a regression or classification model, but we focus on classification in our experiments and discussion.

timate a per-transcription factor, per-sequence region causal effect.

6. Estimate overall per-transcription factor causal effects.

This leaves us with estimates of local (transcription factor and sequence level) and global (transcription factor level) causal effects.

3.2. Exposure and Outcome Effect Size & Standard Error Estimation

As part of the above, we need the predicted difference in both the exposure and outcome value for every mutated, reference sequence pair. Saturation mutagenesis and MC-dropout together provide us with predicted exposure and outcome differences for each mutated sequence, reference sequence pair, but only estimates for the standard errors of the individual predictions not the difference between the two. To obtain the latter quantity, the variance of the differences between the predicted exposure and outcome values for the mutated and reference sequences, we apply the following well-known identity

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) = 2 \cdot \text{Cov}(X, Y). \quad (1)$$

This provides us with a standard error value which we give to Mendelian randomization along with our effect size estimates.

4. Overall Causal Effect Estimation

To estimate overall causal effects at the per-exposure level, we used an inverse-variance weighted random effects meta-analysis.

TODO: Question for David - what to say here?

5. Experimental Results

To test our method, we used a pre-trained DeepSEA model provided by the Kipoi library to estimate the learned causal effect of 36 transcription factors on chromatin accessibility in the HepG2 cell type. We drew our sequence regions from DeepSEA's held-out test set, which was generated via processing the results of ChIP-seq (for transcription factors) and DNase-seq experiments as part of ENCODE project.

For each transcription factor, we randomly sampled 25 (1000 base pair) sequences on which binding was experimentally observed to occur and followed the process described above.

Causal effect estimates vary significantly across transcription factors

The results of our final meta-analysis step, shown in table 1, mply significant variation in the strength of causal relationships between different transcription factors and chromatin accessibility. While all causal effects are positive, certain transcription factors' binding seems to have a very large positive influence on chromatin accessibility. We intend to try and understand the degree to which this reflect modeling assumptions and matches experimental evidence in future work.

Table 1. Per-transcription factor causal effect estimates output by the final step of our algorithm. Columns 1 & 2 contain estimates of the mean effect and its standard error. Column 3, τ^2, contains estimates of variance in the mean produced by heterogeneity in the sequence-level causal effects.

	Treatment Effect		Heterogeneity	
TF	Mean	Std	I^2	τ^2
ATF3	5.682	0.304	1	2.281
BHLHE40	1.753	0.124	1	0.382
CEBPB	0.404	0.019	1	0.009
CEBPD	1.933	0.114	1	0.324
CTCF	0.839	0.066	1	0.11
ELF1	2.116	0.103	1	0.265
EZH2	1.714	0.06	1	0.086
FOSL2	1.454	0.045	1	0.05
FOXA1	2.22	0.055	1	0.073
FOXA2	0.628	0.04	1	0.04
HDAC2	2.579	0.165	1	0.671
HNF4A	2.446	0.127	1	0.402
HNF4G	2.13	0.078	0.999	0.141
MBD4	7.513	0.368	1	3.326
MYBL2	1.704	0.18	1	0.808
NFIC	1.612	0.14	1	0.489
RXRA	2.328	0.111	1	0.309
SP1	1.411	0.072	1	0.126
SP2	8.927	0.114	1	0.31
SRF	15.049	0.544	1	7.346
TAF1	1.359	0.115	1	0.333
TCF12	17.571	0.903	1	20.252
TEAD4	1.654	0.102	1	0.259

Sequence-level causal effect estimates vary significantly for individual transcription factors

6. Discussion

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference* *on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.