Causal Inference & Deep Learning

Notes on Potential Outcomes and Style Transfer

MIT IAP 2018 Compiled by Max W. Shen

Notes are heavily derived from references.

Shen et al. at NIPS 2017 *(Note: The author is a different Shen, not me)* provided an interesting take on the task of deep style transfer with strong parallels to the potential-outcomes framework. Though the work does not directly regard causality with deep learning, a brief discussion could be illuminating for some readers to draw further connections between state-of-the-art machine learning work and ideas in causality.

1 Problem Setup

Consider the following data generating process for text sentences x:

1. A latent style variable y is generated from some distribution $p(y)$; for example, $y \in \{\pm 1\}$ could indicate positive or negative sentiment.
2. A latent content variable z is generated from some distribution $p(z)$; for example, $z \in \mathbb{R}^d$ is a real vector describing word topics.
3. A text sentence x is generated from the conditional distribution $p(x|y, z)$.

Assume access to two observed datasets: $X_1 = \{x_1^1, ..., x_1^n\}$ drawn from $p(x_1|y_1)$ and $X_2 = \{x_2^1, ..., x_2^n\}$ drawn from $p(x_2|y_2)$. Further, assume that X_1 and X_2 are drawn from the same content distribution $p(z)$.

An example of such a dataset may be Yelp reviews for a select number of restaurants. Reviews for the same restaurant can be thought of as sharing the same latent content variable, but some of these reviews may have positive sentiment or style and other reviews may have negative sentiment or style.

The task of style transfer is:

> Given a text sentence x_i with style y_i, what would it look like with the other style y_j?

Concretely, a model that can perform style transfer would be able to "translate" Yelp reviews with negative sentiment into positive sentiment or vice versa without changing its content.

Note how this task parallels the task of estimating counterfactuals in a clinical setting under the standard potential-outcomes framework:

> Given a patient x_i that received treatment y_i, what would have happened if the patient received the other treatment y_j?

2 Deep Style Transfer with a Variational Auto-Encoder

Shen, et al. (2017) consider three approaches for solving the task of deep style transfer, the first of which is a standard variational auto-encoder.

This approach proposes a two-step style transfer process to transfer x_1 to x_2:

1. Infer the content $z \sim p(z|x_1, y_1)$ from x_1.
2. Generate the transferred version $x_2 \sim p(x_2|z, y_2)$ using the inferred content z.

The first step can be accomplished with an encoder $E : X \times Y \to Z$, while the second step can be accomplished with a generator or decoder $G : Y \times Z \to X$.

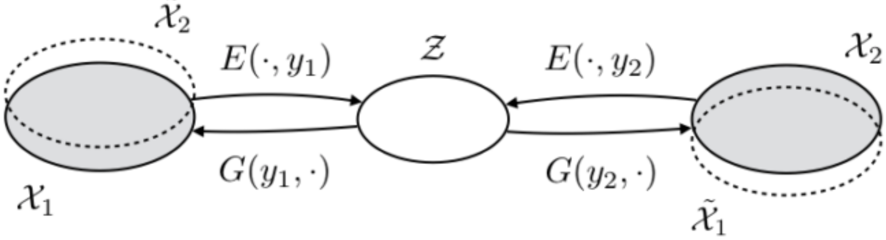

Figure 1: Overview of Proposed Style Transfer Framework. $\tilde{X}_1$ and $\tilde{X}_2$ associated with dotted circles are style-transferred versions of original data.

In a variational auto-encoder, both E and G are implemented using deep neural networks and trained using reconstruction loss on the same style:

$$L_{rec}(\theta_E, \theta_G) = \mathbb{E}_{x_1 \sim X_1}[-\log p_G(x_1|y_1, E(x_1, y_1))] + \mathbb{E}_{x_2 \sim X_2}[-\log p_G(x_2|y_2, E(x_2, y_2))], \tag{1}$$

where θ_E and θ_G are the parameters of the deep neural networks.

In line with the assumption from the generative process, the variational auto-encoder also forces the distribution $p(z)$ to be identical and shared between x_1 and y_1. In standard VAEs, a prior density for $p(z)$ is asserted such as $N(0, I)$, and the following loss function forces $p(z)$ to be shared:

$$L_{KL}(\theta_E) = \mathbb{E}_{x_1 \sim X_1}[D_{KL}(p_E(z|x_1, y_1)||p(z)] + \mathbb{E}_{x_2 \sim X_2}[D_{KL}(p_E(z|x_2, y_2)||p(z)]. \tag{2}$$

The overall objective is $L_{rec} + L_{KL}$.

3 Deep Style Transfer with an Aligned Auto-Encoder

Shen, et al. (2017) argue for various reasons that style transfer may work better when $p(z)$ is not constrained to be a simple prior density such as $N(0, I)$. They propose the relaxed objective

$$\theta^* = \arg\min_{\theta} L_{rec}(\theta_E, \theta_G)$$
$$\text{s.t. } E(x_1, y_1) \text{ shares the same distribution with } E(x_2, y_2). \tag{3}$$

In the aligned auto-encoder, they propose using an adversarial neural network D to enforce $E(x_1, y_1)$ to share the same distribution with $E(x_2, y_2)$ with the loss function

$$L_{adv}(\theta_E, \theta_D) = \mathbb{E}_{x_1 \sim X_1}[-\log D(E(x_1, y_1))] + \mathbb{E}_{x_2 \sim X_2}[-\log(1 - D(E(x_2, y_2)))]. \tag{4}$$

In such a setting, the overall objective is a min-max game between the encoder E, generator G, and discriminator D:

$$\min_{E,G} \max_{D} L_{rec} - \lambda L_{adv}. \tag{5}$$

4 Deep Style Transfer with a Cross-Aligned Auto-Encoder

Building off the idea of an aligned auto-encoder, Shen, et al. (2017) proposes a cross-aligned auto-encoder which uses two adversaries instead of just one.

By assuming that $p(z)$ is shared between X_1 and X_2, the generative process also assumes that transferred X_1 should be distributionally equivalent to X_2 and vice versa. Therefore, two adversaries are used: D_1 which distinguishes between real x_1 and transferred x_2, and D_2 which distinguishes between real x_2 and transferred x_1.

For further details on their work and results, please refer to the paper.

5 References

Shen, T., Lei, T., Barzilay, R., & Jaakkola, T. (2017). Style Transfer from Non-Parallel Text by Cross-Alignment, (NIPS), 112. Retrieved from http://arxiv.org/abs/1705.09655