TODO

Kenny Chiu

September 21, 2021

1 Conceptual summary

The paper by Mouli and Ribeiro [MR21] examines the problem of extrapolating patterns learned from training data collected from a single environment to data collected from other environments. This problem context falls under the idea of *domain adaptation* that has been explored in recent literature [Far+20]. However, a key assumption in Mouli and Ribeiro's work that distinguishes it from previous work in the literature is that the training data come from a single environment as opposed to multiple environments. Several previously proposed methods for domain adaptation—such as *Invariant Risk Minimization* [Arj+20] (IRM)—rely on training data from multiple environments and therefore would fail under this problem context. Mouli and Ribeiro take a different approach by viewing extrapolation as counterfactual reasoning in a specified structural causal model (SCM) and assuming known (linear automorphism) group structures on the non-causal mechanisms. Under this formulation, Mouli and Ribeiro introduce a learning framework for the single-environment context that is able to learn invariances that do not contradict the data. In this conceptual summary, we review the key contributions of the paper by Mouli and Ribeiro [MR21] and discuss the strengths and weaknesses of their approach.

1.1 Key differences from previous work

Various methods for domain adaptation have been proposed in the literature, and how the work by Mouli and Ribeiro [MR21] relates to these methods are highlighted in their paper. For example, methods based on causal inference such as IRM and *Independent Causal Mechanisms* [Par+18] (ICM) broadly involve learning some internal representation of the data that is invariant to environment non-causal mechanisms. The invariant representation is learned from the training data which come from multiple environments. When the data come from a single environment, the representation cannot determine which aspects of the data are environment-specific and so the representation is unlikely to extrapolate to new environments. The learning framework proposed by Mouli and Ribeiro works with single-environment data and has an advantage over existing methods in these settings.

Another well-known approach to domain adaptation is based on data augmentation [CDL20] where training is done with not only the original data but also proper transformations of the data. Mouli and Ribeiro explains that data augmentation is a type of *forced group invariance* (i.e., forced *G-invariance*) where certain transformations of the data may actually introduce contradictions (e.g., trying to enforce rotation invariance in images of digits, but digits 6 and 9 are not invariant to 180^o rotations). Mouli and Ribeiro's proposed learning framework aims to learn only the invariances that do not contradict the training data.

1.2 Main contributions

The main contributions of the paper by Mouli and Ribeiro [MR21] include a formulation of the single-environment extrapolation problem, a learning framework based on the formulation that aims to learn the non-contradicting invariances, and an empirical evaluation of standard neural networks versus neural networks trained using the proposed learning framework.

Mouli and Ribeiro's formulation of the single-environment extrapolation problem is based on the ICM literature where a SCM is used to describe the input variables that influence or are irrelevant to the internal representation across environments (TODOcite?). Extrapolation is then seen as counterfactual reasoning where being able to extrapolate to different environments is tied to the representation being invariant to interventions on non-causal environment variables. Mouli and Ribeiro extends this idea by assuming known linear automorphism groups acting on the non-causal variables, in which case extrapolation of a representation is equivalent to the representation being counterfactually invariant to a group (i.e., *CG-invariant*). This extension is the crux of the formulation that allows the proposed learning framework to work with single-environment data.

The learning framework proposed by Mouli and Ribeiro aims to learn an internal representation that is CG-invariant to specified groups that do not contradict the training data. Mouli and Ribeiro show that CG-invariance is stronger than G-invariance (Theorem 1), but when the group acting on the non-causal variables is a normal subgroup of the group acting on all variables, then G-invariance also implies CG-invariance (Theorem 2). These results establish the group conditions under which it is sufficient for the model to learn G-invariances in place of CG-invariances.

The challenge in learning the G-invariances that do not contradict the training data is due to the fact that the set of non-causal variables among all variables is unknown. To learn the invariances for the unknown set, Mouli and Ribeiro require the groups to be (finite) linear automorphisms. A group-invariant transformation can be constructed by averaging over members of the group (the *Reynolds operator*, Lemma 1). For linear transformations, the averaged transformation is a projection operator with eigenvalues 1 and 0. The left eigenspace spanned by eigenvectors with eigenvalue 1 represents the space of transformations that are invariant to the group (Lemma 2). Mouli and Ribeiro exploit this property by computing the intersection of these subspaces for all subsets of the set of groups, and the set of invariant subspaces can then be partially ordered by the size of their corresponding subset (i.e., the *strength* of the invariance, Theorem 3). The objective optimized in the learning framework then includes a regularization term that encourages learning a representation with the strongest G-invariance that does not significantly contradict the data. The key aspects of this learning framework include needing to specify known groups, requiring the groups to be linear automorphisms and, in doing so, being able to automatically learn the G-invariances that do not (significantly) contradict the data.

Mouli and Ribeiro evaluated neural networks trained using their proposed learning framework on various image tasks and array tasks. Their results broadly suggest that

1. standard neural networks do well when interpolating but not when extrapolating,
2. neural networks trained with forced G-invariances do poorly when interpolating but do well when extrapolating, and
3. neural networks trained with their learning framework generally do well when interpolating and when extrapolating.

1.3 Limitations

The main limitations of the learning framework proposed by Mouli and Ribeiro [MR21] originate from the assumptions made. To allow for single-environment data, the framework requires that the invariance groups acting on the data are known and specified. Furthermore, to enable automatic learning of invariances that do not contradict the training data, the groups are also restricted to be linear automorphisms. This framework cannot be used if no linear group structure could placed on the transformations that act on the data.

These limitations naturally point to directions for future work where non-linear groups or even non-groups of transformations are considered. It is also worth investigating if a method for single-environment extrapolation that does not require known groups is possible.

2 Technical summary

3 Mini-proposals

3.1 Proposal 1: MY PROPOSAL TITLE

3.2 Proposal 2: MY OTHER PROPOSAL TITLE

4 Project report

References

[Arj+20] M. Arjovsky et al. *Invariant Risk Minimization*. 2020. arXiv: 1907.02893 [stat.ML].

[CDL20] S. Chen, E. Dobriban, and J. Lee. "A Group-Theoretic Framework for Data Augmentation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 21321–21333. URL: https://proceedings.neurips.cc/paper/2020/file/f4573fc71c731d5c362f0d7860945b88-Paper.pdf.

[Far+20] A. Farahani et al. *A Brief Review of Domain Adaptation*. 2020. arXiv: 2010.03978 [cs.LG].

[MR21] S. C. Mouli and B. Ribeiro. *Neural Networks for Learning Counterfactual G-Invariances from Single Environments*. 2021. arXiv: 2104.10105 [cs.LG].

[Par+18] G. Parascandolo et al. *Learning Independent Causal Mechanisms*. 2018. arXiv: 1712.00961 [cs.LG].