Information Transfer in Multitask Learning, Data Augmentation, and Beyond

Hongyang R. Zhang

Northeastern University, Boston, Massachusetts
ho.zhang@northeastern.edu

Abstract. A hallmark of human intelligence is that we continue to learn new information and then extrapolate the learned information onto new tasks and domains (see, e.g., Thrun and Pratt (1998)). While this is a fairly intuitive observation, formulating such ideas has proved to be a challenging research problem and continues to inspire new studies. Recently, there has been increasing interest in AI/ML about building models that generalize across tasks, even when they have some form of distribution shifts. How can we ground this research in a solid framework to develop principled methods for better practice? This talk will present my recent works addressing this research question. My talk will involve three parts, divided as follows.

Vignette I: Revisiting multitask learning from the lens of deep learning theory. Multitask learning is a classical idea that dates back at least to the work of Caruana (1997). In recent years, various approaches to multitask learning based on deepnets have shown great results in many areas. Meanwhile, practitioners have also observed problematic outcomes, where the performances of certain tasks have decreased due to task interference. This problem is known as negative transfer in multitask learning (Rosenstein et al. 2005). How would we diagnose the root cause of negative transfer then? In this talk, we revisit this question from the lens of deep learning theory, which provides a suite of models and techniques to reason about the behavior of deepnets.

Implicit regularization in multitask learning: One of the findings from theoretical studies of deep nets is that even though deepnets are heavily overparametrized, they can still generalize well following implicit regularization. In my recent work at ICLR (2020), we ask: is the same implicit regularization mechanism still at work in multitask learning? Surprisingly, we find that the answer depends on the model capacity of the multitask network: if the shared module's capacity is too large, there is no interference between tasks; if it is too small, there can be destructive interference. Then, we show how to determine interference by a fine-grained notion called task covariance to measure task alignment.

Transfer analysis via random matrix theory: It turns out that rigorously studying the above notion of task covariance requires new tools beyond existing generalization theory techniques. In a followup work on arXiv (2021), we bring in tools from recent developments in random matrix theory, inspired by its recent applications for studying interpolation.

Vignette II: Designing principled methods for robust transfer via task modeling. Could our study also suggest new methods for multitask learning? From an empirical perspective, many studies have designed novel architectures for circumventing negative interference. But one takeaway from our theoretical study is that fitting multiple heterogeneous data distributions into a single model is inherently difficult. Next, we revisit ensemble methods for multitask learning.

Task modeling for cross-task transfer: In a recent work presented at NeurIPS Workshop (2022), we explore surrogate models for fitting multitask predictions. Let S be a subset of source tasks from a set of k source tasks. Our approach estimates a surrogate model to approximate the prediction loss of combining S and a target task t, denoted as $f_t(S)$. If S is similar to t, $f_t(S)$ will be small; otherwise $f_t(S)$ will be large. Thus, extrapolating such multitask scores provides a way to model transfer from source tasks to target. Then, we apply surrogate models to several MTL applications.

Vignette III: Algorithms for data augmentation. Lastly, I will consider a slightly different perspective on data augmentation. While data augmentation is most known as regularization or sample-generating techniques, one may cast it in the framework of multitask learning: each transformation can be viewed as one task. Taking this view, we can then ask about the generalization effects of data augmentation in an overparametrized model. Moreover, we can also use multitask learning algorithms to automate data augmentation.

Conclusion: There are many tantalizing questions following the research described above. Would the generalization theories and algorithms apply to other domains, such as robotics, speech, and RL? Our study only talks about the simplest tasks in DL, but there is sufficient overlap between deepnets used by different communities, so it is conceivable that our methodologies might help other communities.

References

Caruana, R. 1997. Multitask learning. *Machine learning*.

Rosenstein, M. T.; Marx, Z.; Kaelbling, L. P.; and Dietterich, T. G. 2005. To Transfer or Not To Transfer.