Towards Near-optimal, Scalable and Practical Anytime Predictors

Abstract:

Machine learning predictors can often be tuned to optimize performance at a fixed test-time budget. However, real world applications often have varying test-time budget. E.g., as an autonomous vehicle moves faster, it needs more frequent obstacle detection; as more requests are sent to a server, it needs to process faster to maintain low request latency. One approach to optimize predictions for any test-time budget is via an anytime algorithm, an algorithm that produces a prediction quickly and continuously refines the result until the budget runs out. Such an algorithm can dynamically adjust to and efficiently use any test-time budget. This work advances anytime prediction algorithms theoretically and practically, from proving near-optimal guarantees in certain settings to developing scalable and competitive anytime predictors for real-world applications in object recognition and scene parsing.

We first propose simple and fast-to-implement greedy methods for the common machine learning setting where the prediction is a generalized linear model and the computational cost is dominated by feature computation. The proposed greedy ordering of feature computation provably leads to near-optimal performance at each anytime prediction. Additionally, if one pays a constant fraction of extra cost, we further prove that we have near-optimal predictions at any budget.

Noting that many anytime algorithms are meta-algorithms that combine multiple predictors using variants of boosting, we next develop streaming gradient boosting to scale up boosting for large data-sets. If each streaming weak learner can fit its target better than noise, the proposed streaming boosting algorithms provably achieve no regret against any predictors under strongly convex loss.

Recognizing that many state-of-the-art predictions are from complex and expensive neural networks, we develop anytime neural networks for practical and competitive anytime predictions at the state-of-the-art level. We add auxiliary predictions to existing networks, and optimize auxiliary losses jointly in an oscillating weighted sum. The resulting anytime predictions from these auxiliary predictions are shown to be competitive against the state-of-the-art at each budget. With a constant fraction of extra cost, we can further assemble a sequence of exponentially deepening anytime networks to achieve frequent and near-optimal anytime predictions. Finally, we apply anytime augmentations on multiple networks, and develop real-world perception applications on visual recognition and semantic segmentation on challenging data-sets, such as ILSVRC and MSCOCO. We also propose to guide the anytime predictions using model distillation and compression techniques on these large data-sets to transfer knowledge of existing complex models into the early anytime predictions.

Speaker:

Hanzhang Hu

Thesis Committee:

Drew Bagnell (Co-Chair)
Martial Hebert (Co-Chair)
Ruslan Salakhutdinov
Rich Caruana (Microsoft Research)

Aug. 30, 2017
4:00pm
8102 GHC

Link to draft document:
https://www.dropbox.com/s/qag5fm2xmyue93a/Hanzhang_proposal_anytime_prediction.pdf?dl=0