Hyperparameter Optimization and DeepHyper

Prasanna Balaprakash and Misha Salim
Argonne National Laboratory
Supervised (deep) learning

- Does it do well on the training data?
  - NO
  - Increase model complexity

- Does it do well on the testing data?
  - NO
  - Collect more data

Magic/Art (60-70%)

Ng A., Challenges of Deep Learning, GTC, 2015
Automated machine learning for deep learning

Lower-level problem:
\[
\text{solve } \min_w \text{ err}_T ([X_A, X_P]; T; w)
\]

Upper-level problem:
\[
\text{solve } \min_{X_A, X_P} \text{ err}_V ([X_A, X_P]; V; w^* [X_A, X_P])
\]

Architecture space Hyperparameter space
DeepHyper: Scalable AutoML package

Hyperparameter search
- AMBS, Hyperband, DEAP, etc

Neural architecture search
- RL with A2C, A3C, random

Workflow
- Balsam

https://github.com/deephyper/deephyper
AMBS: Asynchronous model-based search

- Framework:
  - Initialization phase
    - Random or Latin hypercube sampling
  - Iterative phase
    - Fit model
    - Sample using the model

Example Surrogate Model Fitted to Sampled Performance (iterative refinement improves the learning model)
Bayesian optimization

- Usual Gaussian process regression cannot handle nonordinal space natively
- Appropriate methods: random forest, extra tree regressor, Bayesian NN
- We use Random Forest
Bayesian optimization

\[ LCB(x, \beta) = \mu(x) - \beta \times \sigma(x) \]
Bayesian optimization

LCB \((x) = \mu(x) \times \)
Bayesian optimization

\[ \text{LCB}(x) = \mu(x) \times \text{LCB} \]

Where:
- \( \mu(x) \) is the mean function,
- \( \text{LCB} \) is the Lower Confidence Bound.
Bayesian optimization

\[ \text{LCB}(x) = \mu(x) \times (x) \]
Bayesian optimization

LCB (x) = \mu(x) \cdot \mathcal{I}(x)

- True (unknown)
- \mu(x)
- Observations

- LCB(x)
- Next sample point
Multipoint asynchronous sampling

Naive

Conditioned
Constant liar scheme for asynchronous update

\[ LCB_k(x, \beta) = \mu(x) - \beta \times \sigma(x) \]

\[ f(\hat{x}_{LCB_k}) = \mu(\hat{x}_{LCB_k}) \]

\[ LCB_{k+1}(x, \beta) = \mu'(x) - \beta \times \sigma'(x) \]

\[ f(\hat{x}_{LCB_{k+1}}) = \mu'(\hat{x}_{LCB_{k+1}}) \]
Comparison of search methods

Target platform: Theta@ALCF (128 KNL nodes)
Stopping criterion: 2 hours
Comparison of search methods

Target platform: Theta@ALCF (128 KNL nodes)
Stopping criterion: 2 hours
Scaling search methods

Target platform: Theta@ALCF (128 KNL nodes)
Benchmark: rnn2; Stopping criterion: 2 hours
Target platform: Cooley (64 nodes Haswell + NVIDIA Tesla K80)
Benchmark: cifar10cnn; Stopping criterion: 1 hour
DeepHyper: Scalable AutoML package

- Hyperparameter search: AMBS, Hyperband, DEAP, etc
- Neural architecture search: RL with A2C, A3C, random
- Workflow: Balsam

https://github.com/deephyper/deephyper
Acknowledgements

DOE Early Career Research Program, ASCR

CANDLE ECP Project

Argonne Leadership Computing Facility

Laboratory Directed Research and Development (LDRD)