JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

CLSP Seminar

Feb 28. 2020

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Speaker 1

Xuan Zhang

Hyperparameter Optimization of Neural Machine Translation Systems

Xuan Zhang, Kevin Duh

Overview

0. Introduction to Hyperparameter Optimization (HPO)

I. A new benchmarking dataset for HPO of NMT systems

II. A new HPO method

Overview

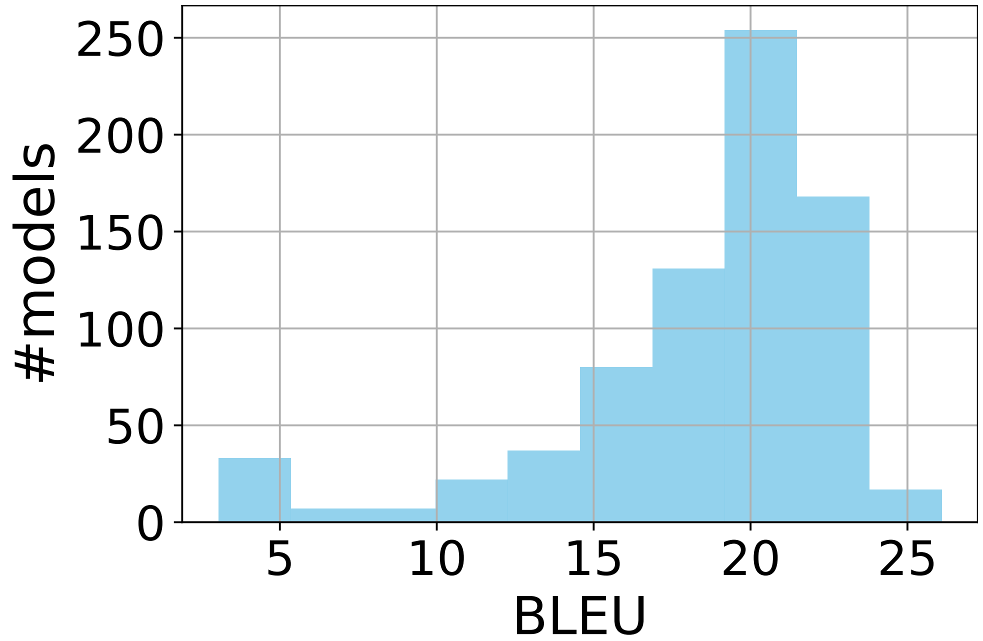
0. Introduction to Hyperparameter Optimization (HPO)

I. A new benchmarking dataset for HPO of NMT systems

II. A new HPO method

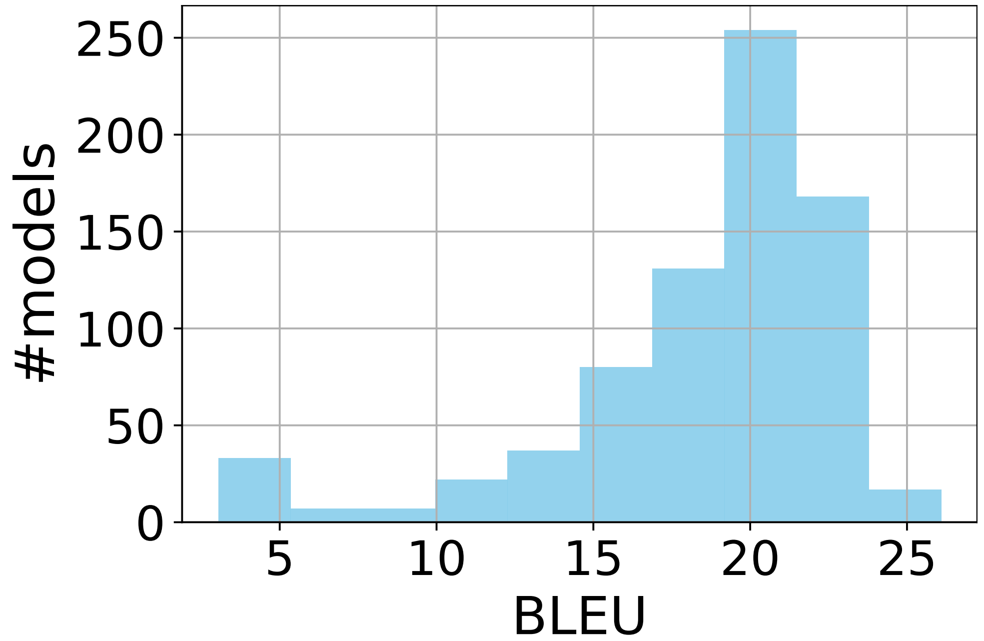
0. Hyperparameter Optimization - Motivation

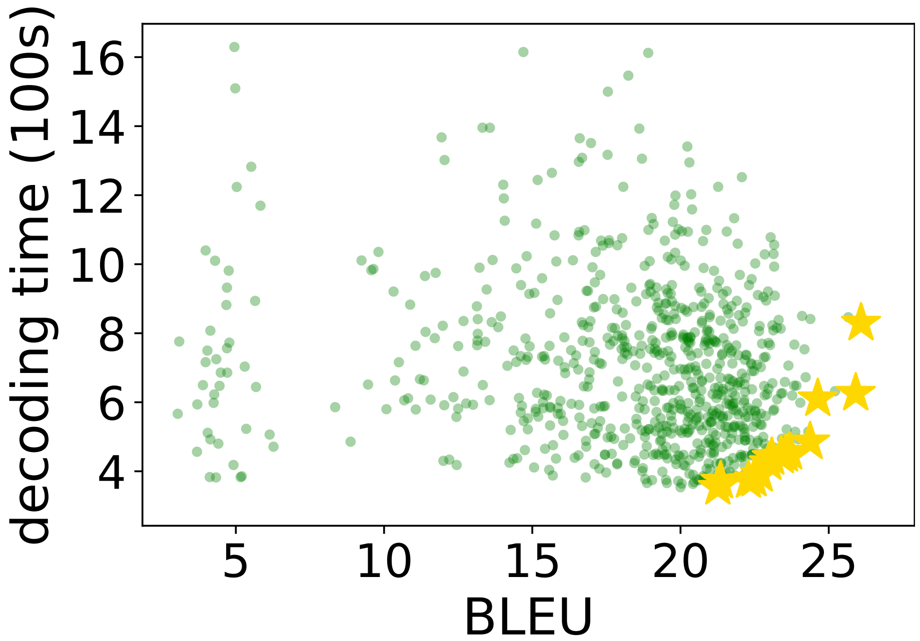
- Choosing effective hyperparameters is crucial for building strong NMT systems: *initial learning rate, batch size, etc.*



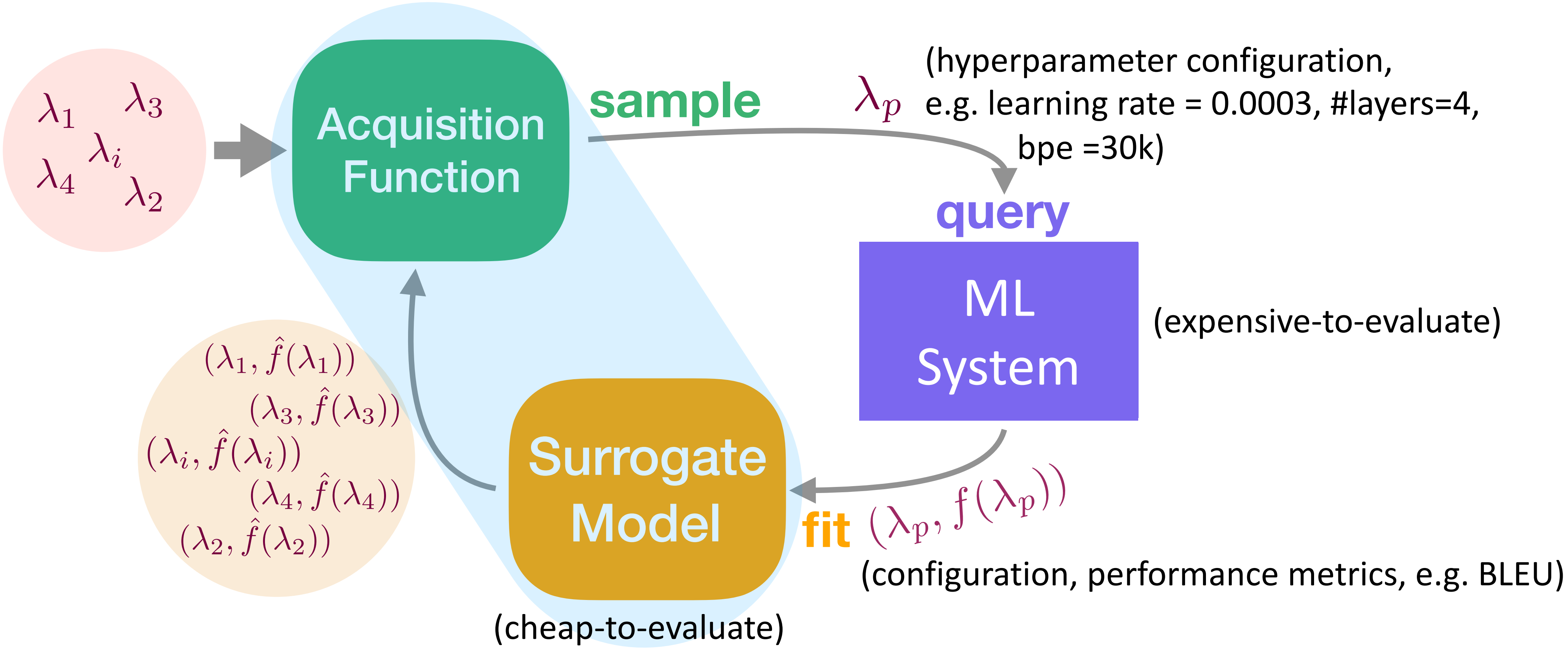
0. Hyperparameter Optimization - Motivation

- Choosing effective hyperparameters is crucial for building strong NMT systems: *initial learning rate, batch size, etc.*
- Manual tuning is tedious and unreliable, especially when optimizing multiple objectives: *BLEU & decoding time.*





0. Hyperparameter Optimization - Approach



Goal: find a hyperparameter configuration $\boldsymbol{\lambda}_\star = \arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} f(\boldsymbol{\lambda})$ with as few evaluations of $f(\cdot)$ as possible.

0. Hyperparameter Optimization - Challenges

- State-of-the-art NMT models require significant **computational resources** for training.

- NMT models usually have large hyperparameter **search space**.

-> It is expensive to compare HPO methods on NMT tasks.

Overview

0. Introduction to Hyperparameter Optimization (HPO)

I. A new benchmarking dataset for HPO of NMT systems

II. A new HPO method

I. Table-Lookup Datasets for NMT HPO

- First, we pretrain a large number of NMT systems covering a wide range of hyperparameter configurations, and record their performance metrics.
- Then, we constrain our HPO methods to sample from this finite set of models.

I. Table-Lookup Datasets for NMT HPO

- **MT Data:**
  Ted talks: **zh-en, ru-en**; WMT: **ja-en, en-ja**; low-resource: **sw-en, so-en**

- **Model:** Transformer

- **Hyperparameters:**
  preprocessing configurations: bpe
  training settings: initial learning rate
  architecture designs: #layers, embed, #hidden, #heads in self-attention

- **Objectives:**
  translation accuracy: dev BLEU, dev perplexity
  computational cost: decoding time, training time, GPU memory for training, total number of model parameters

I. Table-Lookup Datasets for NMT HPO

dataset	**bpe** (1k)	**#layers**	**#embed**	**#hidden**	**#att_heads**	**init_lr** (10^{-4})
zh, ru, ja, en	10, 30, 50	2, 4	256, 512, 1024	1024, 2048	8, 16	3, 6, 10
sw	1, 2, 4, 8, 16, 32	1, 2, 4, 6	256, 512, 1024	1024, 2048	8, 16	3, 6, 10
so	1, 2, 4, 8, 16, 32	1, 2, 4	256, 512, 1024	1024, 2048	8, 16	3, 6, 10

Table 1: Hyperparameter search space for the NMT systems.

Dataset	**#models**	**Best BLEU**	**bpe**	**#layers**	**#embed**	**#hidden**	**#att_heads**	**init_lr**
zh-en	118	14.66	30k	4	512	1024	16	3e-4
ru-en	176	20.23	10k	4	256	2048	8	3e-4
ja-en	150	16.41	30k	4	512	2048	8	3e-4
en-ja	168	20.74	10k	4	1024	2048	8	3e-4
sw-en	767	26.09	1k	2	256	1024	8	6e-4
so-en	604	11.23	8k	2	512	1024	8	3e-4

Table 2: For each language pair, we report the number of NMT systems trained on it, the oracle best BLEU we obtained and its corresponding hyperparameter configuration.

Overview

0. Introduction to Hyperparameter Optimization (HPO)

I. A new benchmarking dataset for HPO of NMT systems

II. A new HPO method

II. Graph-Based Semi-Supervised Learning as a HPO Method

- **Semi-supervised learning:** utilize a handful of labeled data and a large amount of unlabeled data to improve prediction accuracy.

- **Graph-Based Semi-Supervised Learning (Zhu et al., 2003):** describes the structure of data with a graph.

vertex: (data point, label) → (configuration, model performance)
edge weight: similarity between vertices → similarity between configurations
smoothness: neighbors connected by edges tend to have similar labels.
Predict: Labels can propagate throughout the graph.