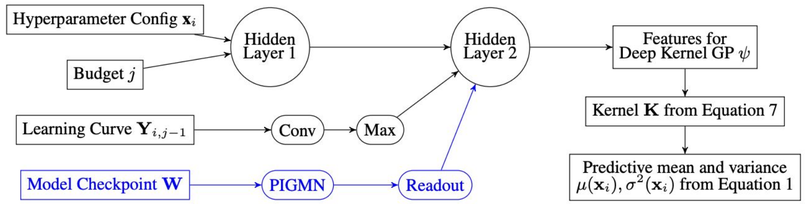
Improving Hyperparameter Optimization with Checkpointed Model Weights

Nikhil Mehta, Jonathan Lorraine, Steve Masson, Ramanathan Arunachalam, Zaid Pervaiz Bhat, James Lucas, Arun George Zachariah

Introduction

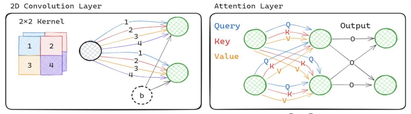
- Hyperparameter Optimization (HPO) is a critical and expensive part of model design.
  - Ex., choosing a foundational model to use
- Classical black-box HPO ignores problem specific structure.
  - Ex., grid/random search
- Gray-box HPO uses problem specific structure to improve performance.
  - Ex., multi-fidelity or gradient-based
- We propose how to **leverage logged weight checkpoints to improve HPO**.
- We make all code available.

Our Method - FMS



- DyHPO[1] is a successful multi-fidelity Bayesian Optimization HPO.
- Forecasting Model Search (FMS) builds on DyHPO, with relevant changes shown in blue.
- We featurize logged weight checkpoints via permutation-invariant graph metanets (PIGMNs)
  - Permutation-invariance enhances data efficiency
- The checkpoints are a rich source of info on architecture, dataset, loss, and optimization.

Our Method - Some Specifics



PIGMN graph construction [2]

$$\mathcal{G}^{(0)}(\mathbf{W})$$

$$\mathcal{G}^{(l+1)} = \sigma\left(\sum_{k=1}^{K} \Theta_k^{(l)} * \mathcal{G}^{(l)}\right)$$

$$\mathbf{K}(\boldsymbol{\theta}, \mathbf{w}) := k(\psi(\mathbf{x}_i, \mathbf{W}_i, \mathbf{Y}_{i,j-1}, j; \mathbf{w}), \psi(\mathbf{x}_{i'}, \mathbf{W}_{i'}, \mathbf{Y}_{i',j'-1}, j'; \mathbf{w}); \boldsymbol{\theta})$$

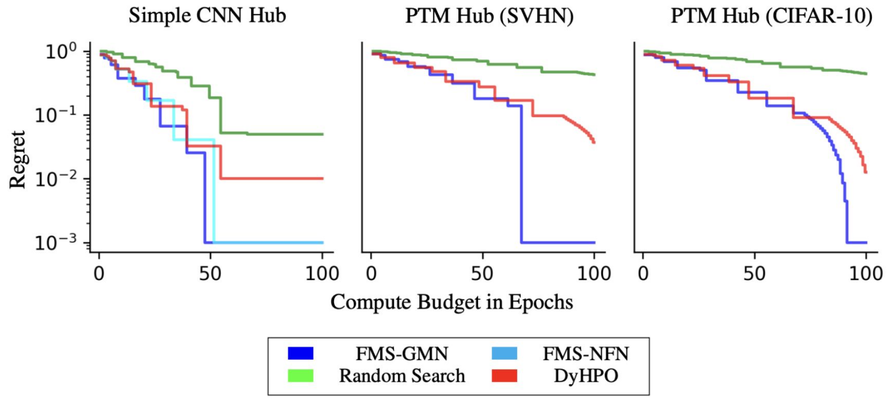
$$\nabla_{\boldsymbol{\theta},\mathbf{w}} \mathcal{L}(\mathcal{D}) = -\left(\mathbf{y}^\top \mathbf{K}(\boldsymbol{\theta}, \mathbf{w}, \mathcal{D})^{-1} \mathbf{y} - \mathrm{Tr}\left(\mathbf{K}(\boldsymbol{\theta}, \mathbf{w}, \mathcal{D})^{-1}\right)\right)$$

Algorithm 2 Our Forecasting Model Search (FMS) method, with changes from DyHPO in blue.

1: Initialize $\mathcal{D}$ with any preexisting evaluations of configurations $\mathbf{x}$, , losses $\mathbf{Y}_{i,j-1}$, their weights $\mathbf{W}$, and budgets j and update the GP parameters $\boldsymbol{\theta}$ and $\mathbf{w}$ via gradient-based optimization with $\nabla_{\boldsymbol{\theta},\mathbf{w}}\mathcal{L}$ from Equation 4
2: **while** computational budget not exhausted **do**
3: Select $(\mathbf{x}_i, \mathbf{W}_i, j)$ by maximizing $\mathrm{EI}_{\mathrm{MF}}(\mathbf{x}, \mathbf{W}, j|\mathcal{D})$
4: Evaluate configuration and budget: $y_i = f(\mathbf{x}, \mathbf{W}, j)$, with loss trajectory $\mathbf{Y}_{i,j-1}$
5: Update $\mathcal{D}$ with $(\mathbf{x}_i, \mathbf{W}_i, j, \mathbf{Y}_{i,j-1})$, effectively checkpointing the weights $\mathbf{W}$
6: Optimize GP parameters $\boldsymbol{\theta}$ and $\mathbf{w}$ via gradient-based optimization for N steps or until termination criteria are satisfied with $\nabla_{\boldsymbol{\theta},\mathbf{w}}\mathcal{L}$ from Equation 4
7: **return** configuration $\mathbf{x}_i$ with the best observed performance y_i, and its learned parameters $\mathbf{W}_i$

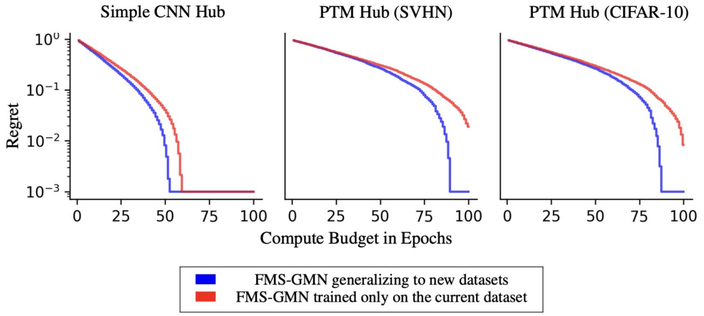
Experimental Results

- We look at selecting a pre-trained (foundational) model from a hub, along with subsequent hyperparameters for fine tuning.



Experimental Results

- FMS exhibits transfer learning, by showing improved performance when trained on checkpoints from other setups.



- Various ablations over FMS vs. DyHPO and design choices for FMS provided in the paper.

Our Vision of the Future

- We envision building "foundational" HPO methods which efficiently tune a broad set of problems by wielding large amounts of optimization metadata.
  - Now we can use the metadata of checkpointed weights

Additional Links

- Code: github.com/NVlabs/forecasting-model-search
- Webpage: research.nvidia.com/labs/toronto-ai/FMS/
- [1] Wistuba, Martin, Arlind Kadra, and Josif Grabocka. "Supervising the multi-fidelity race of hyperparameter configurations." Advances in Neural Information Processing Systems 35 (2022): 13470-13484.
- [2] Lim, Derek, et al. "Graph Metanetworks for Processing Diverse Neural Architectures." The Twelfth International Conference on Learning Representations.