Combination of Hyperband and Bayesian Optimization for Hyperparameter Optimization in Deep Learning

Jiazuho Wang, Jason Xu, Xuejun Wang
Blippar

arXiv:1801.01596v1

Information & Intelligence System Lab.
Sang Heon Lee
lawlee1@naver.com

2018/09/06

Introduction

- Hyperparameter optimization
  - Finding particular hyperparameters that optimizes some evaluation criterion, e.g., loss on a validation set
- Hyperband algorithm
  - Allocate more budget(time) to more promising hyperparameter settings
  - Early-stopping strategy
- Limitation of Hyperband
  - Treats all hyperparameter points as being independent from each other
    - Hyperparameter space is usually smooth

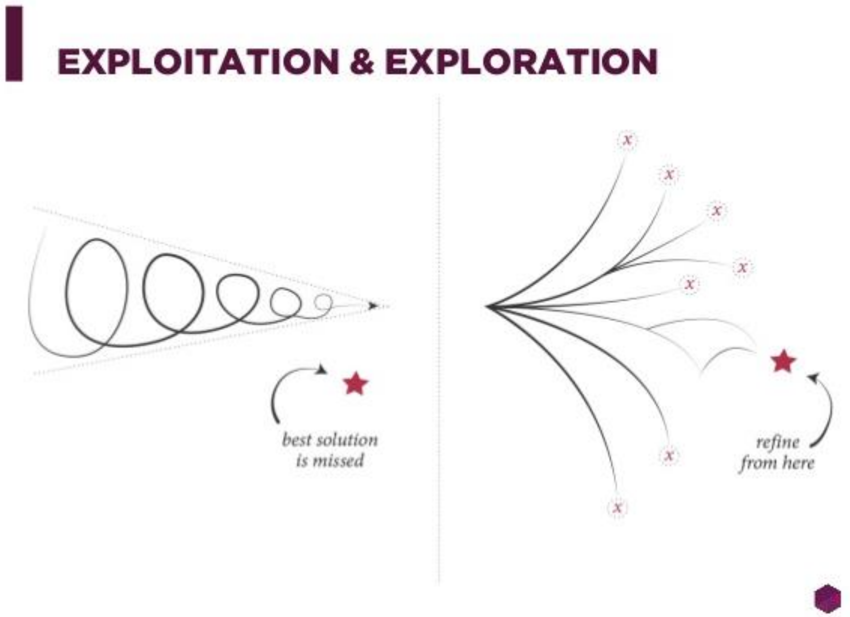
Introduction (Cont'd)

- Bayesian Optimization
  - Utilize the information of previous trial points
  - Balances between exploitation and exploration
  - Allocates computational budget uniformly, which causes efficiency issues

- Propose to combine Hyperband algorithm and Bayesian optimization
  - Fully utilize what we learned through history and focus our attention on really promising ones to avoid wasting times

Background

- Optimization : Exploitation vs. Exploration
  - Exploitation : choose the best point until now
  - Exploration : choose a new point to get more information
  - trade-off relations : the proper control of these two behaviors is core



Background (Cont'd)

- Hyperband Algorithm
  - Allocate more resources to more promising hyperparameter configurations
  - More resources allocate, more you can clearly determine the final loss using intermediate loss

(1) Initialize n trial points

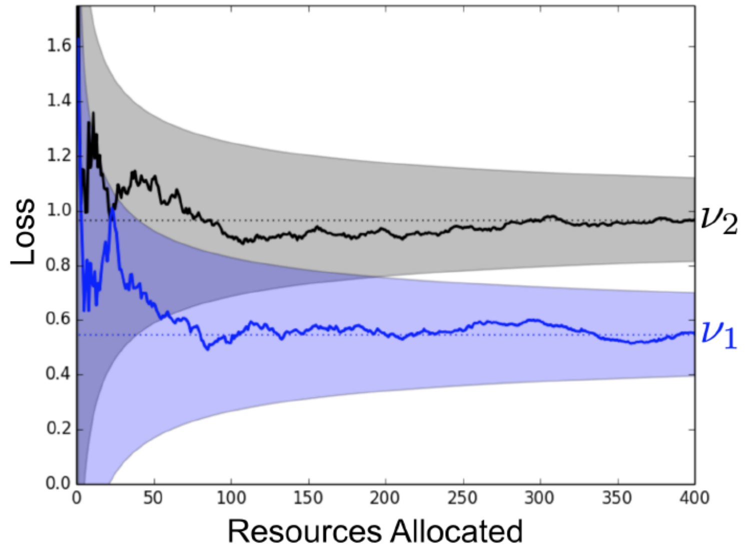
  (2) uniformly allocates a budget to each trial points

  (3) Evaluate each performance

  (4) Remove worse points from the set

(5) Repeat (2), (3), (4) until one point remains

(6) Update n, r and go (1)



Background (Cont'd)

- Hyperband Algorithm
  - R : maximum amount of resource(training time) that can be allocated to a single configuration
  - η : percentage of configurations discarded at each step

Algorithm 1: HYPERBAND algorithm for hyperparameter optimization.

input : R, η (default $\eta = 3$)
initialization : $s_{\max} = \lfloor \log_\eta(R) \rfloor$, $B = (s_{\max}+1)R$
for $s \in \{s_{\max}, s_{\max}-1, \ldots, 0\}$ do
  $n = \lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \rceil, \quad r = R\eta^{-s}$
  // begin SUCCESSIVEHALVING with (n, r) inner loop
  T = get_hyperparameter_configuration(n)
  for $i \in \{0, \ldots, s\}$ do
    $n_i = \lfloor n\eta^{-i} \rfloor$
    $r_i = r\eta^i$
    $L = \{$run_then_return_val_loss$(t, r_i) : t \in T\}$
    T = top_k$(T, L, \lfloor n_i/\eta \rfloor)$
  end
end
return *Configuration with the smallest intermediate loss seen so far.*

[Example] R = 81, η =3

- s_{max} = $\lfloor \log_3 81 \rfloor = 4$
- B = (s_{max}+1)R = 5*81

s=4)

- n = $\left[\frac{5*81}{81}\frac{3\text{^}4}{5}\right]$ = 81, $r = 81 * 3\text{^}(-4) = 1$
- get 81 configurations
- Evaluate, remove with 1/3

s=3)

- n = $\left[\frac{5*81}{81}\frac{3\text{^}3}{4}\right]$ = 34, $r = 81 * 3\text{^}(-3) = 3$
- get 34 configurations
- Evaluate, remove with 1/3

Background (Cont'd)

- Bayesian Optimization
  - Approximate expensive function f using a probabilistic surrogate model

(1) Samples next trial point(x_{t+1}) according to current surrogate model

(2) Evaluate f(x_{t+1})

(3) Update surrogate model based on new data point (x_{t+1}, f(x_{t+1}))

(4) Go to (1) and repeat

Algorithm 2 Bayesian optimization

1:	**initialization:** $D_0 = \emptyset$
2:	**for** $t \in \{1, 2, ...\}$ **do**
3:	$x_{t+1} = argmax_x \boxed{\mu(x\|D_t)}$
4:	Evaluate $f(x_{t+1})$
5:	$D_{t+1} = D_t \cup \{(x_{t+1}, f(x_{t+1}))\}$
6:	Update probabilistic surrogate model using D_{t+1}

acquisition function : Expected Improvement (EI)

$$\mu(x|D_t) = \mathbb{E}(max\{0, f_{t+1}(x) - f(x^+)\}|D_t)$$

Proposed Method

- Limitation of Hyperband and Bayesian Optimization
  - Hyperband : It fails to utilize the **information of history**
  - Bayesian optimization : Each point is allocated **sufficient resources**

- Two methods are perfectly complementary to each other
  - Proposed combination of HB and BO

Proposed Method (Cont'd)

- Combination of Hyperband and Bayesian Optimization

Algorithm 3 Combination of Hyperband and Bayesian optimization

---

input: maximum amount of resource that can be allocated to a single hyperparameter configuration R, and proportion controller η
output: one hyperparameter configuration

initialization: $s_{max} = \lfloor log_{\eta}(R) \rfloor$, $B = (s_{max} + 1)R$
for $s \in \{s_{max}, s_{max} - 1, ..., 0\}$ **do**
    $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil, r = R\eta^{-s}$
    for $i \in 0, ..., s$ **do**
        $n_i = \lfloor n\eta^{-i} \rfloor$
        $r_i = r\eta^i$
        if $i == 0$ **then**
            $X = \emptyset, D_0 = \emptyset$
            for $t \in \{1, 2, ..., n_i\}$ **do**
                $x_{t+1} = argmax_x \mu(x|D_t)$
                $f(x_{t+1})$ = run_then_return_obj_val(x, r_i)
                $X = X \cup \{x_{t+1}\}$
                $D_{t+1} = D_t \cup \{(x_{t+1}, f(x_{t+1}))\}$
                Update probabilistic surrogate model using D_{t+1}
        else
            F = { run_then_return_obj_val(x, r_i) : $x \in X$ }
            X = top_k($X, F, \lfloor n_i/\eta \rfloor$)
return configuration with the best objective function value

Proposed Method (Cont'd)

- Combination of Hyperband and Bayesian Optimization

initialization: $s_{max} = \lfloor log_\eta(R) \rfloor$, $B = (s_{max} + 1)R$
for $s \in \{s_{max}, s_{max} - 1, ..., 0\}$ **do**
    $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil$, $r = R\eta^{-s}$
    X = get_hyperparameter_configuration(n)
    for $i \in 0, ..., s$ **do**
        $n_i = \lfloor n\eta^{-i} \rfloor$
        $r_i = r\eta^i$
        F = { run_then_return_obj_val(x, r_i) : $x \in X$ }
        X = top_k($X, F, \lfloor n_i/\eta \rfloor$)
return configuration with the best objective function value

Hyperband

initialization: $s_{max} = \lfloor log_\eta(R) \rfloor$, $B = (s_{max} + 1)R$
for $s \in \{s_{max}, s_{max} - 1, ..., 0\}$ **do**
    $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil$, $r = R\eta^{-s}$
    for $i \in 0, ..., s$ **do**
        $n_i = \lfloor n\eta^{-i} \rfloor$
        $r_i = r\eta^i$

Bayesian Optimization

        if $i == 0$ **then**
            $X = \emptyset$, $D_0 = \emptyset$
            for $t \in \{1, 2, ..., n_i\}$ **do**
                $x_{t+1} = argmax_x \mu(x|D_t)$
                $f(x_{t+1})$ = run_then_return_obj_val(x, r_i)
                $X = X \cup \{x_{t+1}\}$
                $D_{t+1} = D_t \cup \{(x_{t+1}, f(x_{t+1}))\}$
                Update probabilistic surrogate model using D_{t+1}
        else
            F = { run_then_return_obj_val(x, r_i) : $x \in X$ }
            X = top_k($X, F, \lfloor n_i/\eta \rfloor$)
return configuration with the best objective function value

Proposed Method

Proposed Method (Cont'd)

- Combination of Hyperband and Bayesian Optimization

initialization: $s_{max} = \lfloor log_\eta(R) \rfloor$, $B = (s_{max} + 1)R$
for $s \in \{s_{max}, s_{max} - 1, ..., 0\}$ **do**
  $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil$, $r = R\eta^{-s}$
  for $i \in 0, ..., s$ **do**
    $n_i = \lfloor n\eta^{-i} \rfloor$
    $r_i = r\eta^i$
    if $i == 0$ **then**
      $X = \emptyset$, $D_0 = \emptyset$
      for $t \in \{1, 2, ..., n_i\}$ **do**
        $x_{t+1} = argmax_x \mu(x|D_t)$
        $f(x_{t+1}) = \text{run\_then\_return\_obj\_val}(x, r_i)$
        $X = X \cup \{x_{t+1}\}$
        $D_{t+1} = D_t \cup \{(x_{t+1}, f(x_{t+1}))\}$
        Update probabilistic surrogate model using D_{t+1}
    else
      $F = \{\ \text{run\_then\_return\_obj\_val}(x, r_i) : x \in X\}$
      $X = \text{top\_k}(X, F, \lfloor n_i/\eta \rfloor)$
return configuration with the best objective function value

[Example] R = 27, η =3
- s_{max} = $\lfloor \log_3 27 \rfloor = 3$, B = $(s_{max}+1)$R = 4*81

s=3)
- n = $\left[\frac{4*81}{81}\frac{3^3}{4}\right]$ = 27, $r = 27 * 3^{(-3)} = 1$

① **i=0 : Bayesian Optimization**
- n_0=27, r_0=1
- With BO, 27 configurations (ran with 1 resource)

② **i=1 : Hyperband**
- n_1=9, r_1=3
- Run : 27 configurations + 3 resources
- Extract 9 configurations (ran with 3 resource)

③ **i=2 : Hyperband**
- Run : 9 configurations + 9 resources
- Extract 3 configurations (ran with 9 resources)

④ **i=3 : Hyperband**
- Run : 3 configurations + 27 resources
- Extract 1 configuration (ran with 1 resource)

s=2) ...

Experiments

- Two experiments
  - Choose TPE as the Bayesian optimization part
    - GP : model $p(y|x)$ <-> TPE : model $p(x|y)$ and $p(y)$
    - due to its supreme performance than GP

  - Named as **Hyperband_TPE**
    - 3 baselines : Random search, TPE(BO), Hyperband

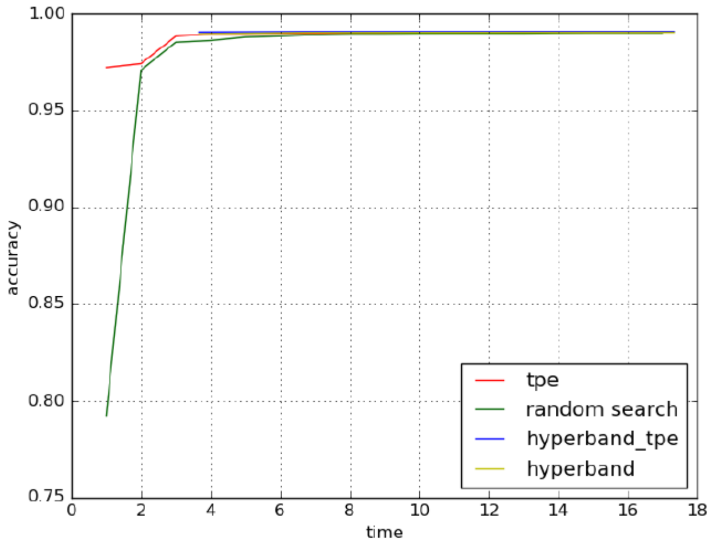
Experiments (Cont'd)

1. Hyperparameter Optimization

1.1. LeNet on MNIST

- 4 hyperparameters : learning rate, batch size, # of filters for 2 conv layers
- Resource : # of training images
  - R = 81, η = 3 ,unit of R : empirically settings

- Result
  - 4 approaches quickly converges
  - simple problems, no sophisticated approach is needed



multiple of R

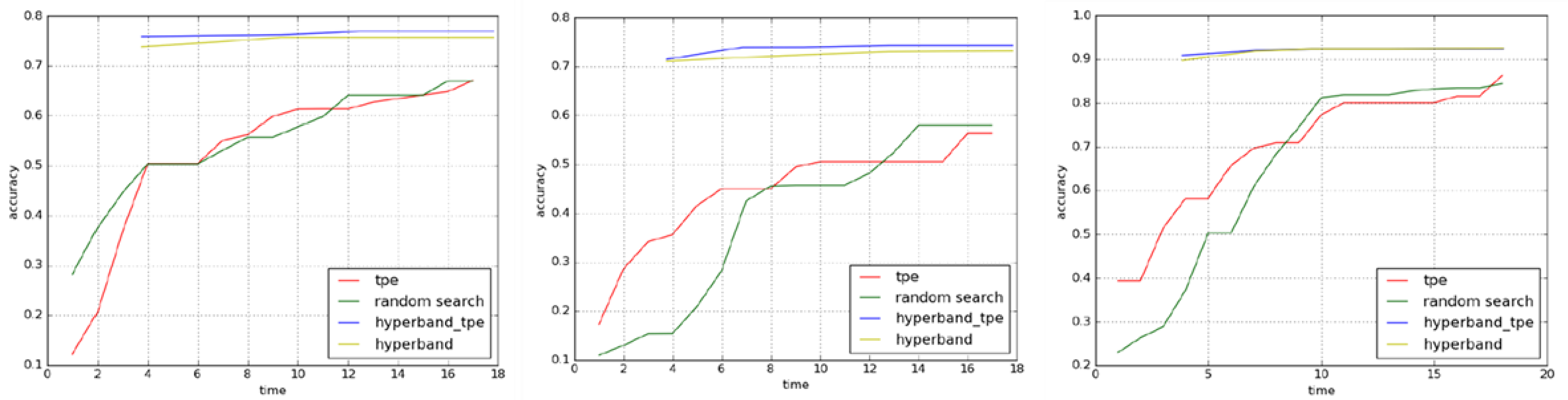
Figure 1: Average accuracy across 10 trials for LeNet on MNIST.

Experiments (Cont'd)

1. Hyperparameter Optimization

1.2. AlexNet on Cifar10, MRBI, and SVHN

- 8 hyperparameters : learning rate, l2 penalty for 4 layers, …
- Resource : # of training images
  - R = 300(Cifar10, MRBI) or 600(SVHN), η = 4, unit of R : empirically settings



multiple of R

(a) Cifar10 (b) MRBI (c) SVHN

Figure 2: Average accuracy across 10 trials for AlexNet on Cifar10, MRBI, and SVHN.

Experiments (Cont'd)

2. Decomposition of SSD

- Find low rank decomposed approximation NN of the Single Shot Detector(SSD)
  - SSD : one of state-of-the-art object detector
  - To balance well between accuracy and speed (on low end GPU)

- Low rank regularization technique via SVD

  ex) 1 conv layer, N filters(d*d), C channels

  → 2 conv layers : (K filters(d*1), C channels) and (N filters(1*d), K channels)
  - K : hyperparameter that controls degree of information compression
    - smaller K, the quicker but less accurate

- Apply above technique on SSD
  - 21 conv layers : 21 of value K : 21 dimensions of hyperparameter settings

Experiments (Cont'd)

2. Decomposition of SSD

- Dataset : PASCAL VOC dataset
- Resource : # of training images
  - R = 2500, η = 5, unit of R : empirically settings
- objective function : $\alpha \times map + fps$
  - map : accuracy
  - fps : speed
  - α : parameter balancing the map and fps

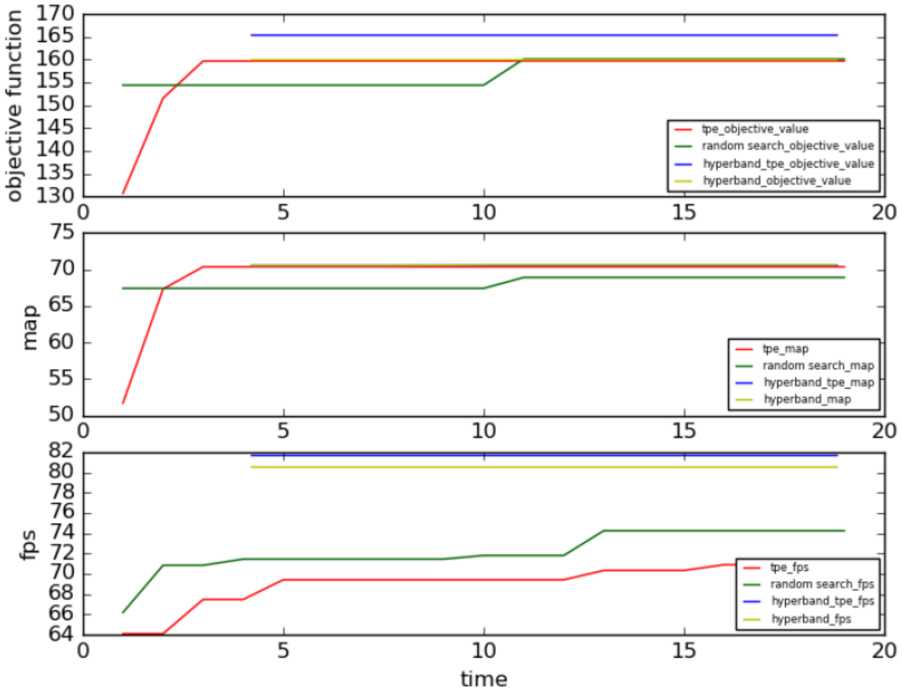


Figure 3: Objective function value, map, and fps for low rank decomposition of SSD.

Conclusion

- Proposed novel approach for hyperparameter optimization
  - Combines the strength of both Hyperband and Bayesian optimization

- Combinational approach could find better hyperparameter more quickly than other approaches
  - Outperforms other approaches by a larger margin, as the problem become more complex and difficult