Literature Review

Christopher MacKinnon

November 2020

1 Hyper-Parameters

Hyper-parameter optimisation is traditionally considered a noisy, black-box optimisation problem. A manual search combing a grid search and expert knowledge has been the standard approach to this problem in the past. It was shown that a random search could perform as well as, or better than, grid search in the same time [7]. This type of search has become a minimum benchmark or starting point for many hyper-parameter optimisation techniques. [14][18] Recently, taking advantage of the increase in parallel and cloud computing power has become a key aspect of these systems, requiring robust methods that can operate efficiently at scale. Hyper-parameters can contain continuous, discrete and categorical variables requiring flexible approaches to find optimal solutions across a wide range of architectures and algorithms.

2 Architecture Search

There has been an increase in interest in Neural Architecture Search (NAS) recently as these methods began to outperform human designed models. [25][23] These systems are able to design at scale creating intricate topologies for large, deep neural networks (DNN), while discovering novel variants in the components used to build these networks. [16] [19]

3 Genetic Algorithms

Genetic algorithms borrow optimisation strategies from those found in nature. One of the main benefits of genetic algorithms is the flexibility in terms of categorical, discrete and continuous variables. GAs are naturally suited to the creation of novel topologies[3] [16], and can often produce targeted architectures that deviate from the generalised templates that commonly used due to their effectiveness across a range of domains. GAs have also been used on a more granular level, creating novel variants of network components suited to a specific task [16]

3.1 Crossover and Genetic Encoding

Crossover is the recombination of two successful networks into a "child" network which is common to many GA approaches to NAS. [21][3][22] This can be implemented through a wide variety of mechanisms based on the network encoding method which is used. Ablation tests on GA systems have shown that while crossover can result in an overall improvement if carried out correctly, it can often be far less significant that other aspects of GA. [22][3]

One of the core components to GA is the genetic encoding of model architectures. This is an important aspect of a successful GA due to the reliance of crossover on this representation of the network to carry meaningful aspects of the models forward. [3] Another issue that is important to consider when designing a genetic encoding methods is the problem of competing conventions or the permutation problem[1][3]. In the context of ANNs this refers to a scenario where the same architecture can be represented by multiple encodings. This can lead to inefficient allocation of computational resources and irregularities during crossover operations. [3]

3.2 NEAT

NeuroEvolution of Augmenting Topologies (NEAT) [3] describes a framework for evolving neural networks which addresses many of the issues faced by GA, such as a formalised system for crossover, genetic encoding, the permutation problem and the protection of innovation. NEATs success is limited to smaller networks, however, there have been several extensions to NEAT, adapting it for larger, deeper networks. [4][16]

NEAT makes use of an innovation number which denotes the order in which mutations occurred to the network. This enables a matching of connections or nodes in the crossover process between networks giving the method a structured process. Speciation is a component of GA which separates models based on genetic distance into subcategories. This protects innovation allowing solutions to be explored and reach maturity without competing with the wider population. In practice this creates a less greedy algorithm supporting a greater diversity of solutions and, in theory, improving performance on multi-modal functions.

Ablation testing was able to show the dependence of the wider system on each of these components. Speciation was shown to offer a very significant contribution to overall success of the method, while aspects such as crossover was shown to have a much smaller, but still statically significant, improvements in performance.

3.3 CoDeepNEAT

CoDeepNEAT [16] is an expansion to NEAT for use with deep neural networks. This approach was able to achieve near parity with SOTA on CIFAR-10 as well as image captioning benchmarks.

CoDeepNEAT is based on a bi-level optimisation approach to the CASH problem, implementing a hierarchical approach in which "blueprints" and modules are evolved independently. A blueprint is a graph of nodes which are replaced with modules to create complete networks for evaluation. The fitness score of a network is applied to both the blueprint and the modules, with the module score being an average of all the networks which contained the blueprint. This method was also applied to LSTM networks, which was able to produce an LSTM variant which outperform the standard LSTM cell.

3.4 N-Level Hierarchical Representation

[19] introduced a nested system for NAS which expands upon the method described in [16] from a bi-level hierarchy to an Nnlevel hierarchy. In this system, primitive operations (i.e. convolutional cells, linear cells) are considered level one. level two representations are a set of graph structures combining these lower level cells. These second level representations are then combined replacing the nodes in a graph to create higher level representation.

3.5 Population Based Training

[14] introduced a form of GA which takes inspiration from bandit problems called Population Based Training (PBT). This was able to outperform human tuned networks on reinforcement learning and imagine classification tasks.

This is a system for hyper-parameter optimisation which utilizes a variant of the successive-halving[10] method used in Hyperband (see Sec 4.1) for early-stopping, A portion of the population abandon their search and copy the structure, hyper-parameters and weights of high performing networks. The hyper-parameters of this new replicated network are then mutated randomly to allow for more thorough examination of lucrative search spaces. [20] introduces a similar system targeted at Neural Architecture Search (NAS). This method of exploiting successful networks has the additional advantage of producing hyper-parameter schedules that vary over the training process.

3.6 LEMONADE

[20] approach the problem of NAS with a multi-objective based method that seeks to retain network function across generations via Lamarckian inheritance. For applications in embedded systems or other computational limited environments, the complexity of the network can be an important factor. This system utilizes a Pareto front in the selection process to select a curve of models based on performance and complexity.

3.7 NSGA-NET

NSGA-Net [22] is a GA based multi-objective optimisation system for NAS which is build upon NSGA-II[2] and incorporates Bayesian Optimisation meth-

ods. This approach was able to achieve parity with SOTA on the CIFAR-10 data-set compared with other NAS solutions (in particular RL based methods) at a significantly lower computational cost using a similar multi-objective method as [20].

This approach encodes a group of nodes (a single computational unit, i.e. convolution, pooling in this particular application) as binary strings, these are describes as phases, which are then stacked to describe an entire model. Mutations are applied through bit-flipping of this encoding. Crossover was achieved through a comparison of binary strings where common bits from both parents are retained and unique bits are chosen at random from either parent.

The final stage of this technique involves applying a Bayesian Network to exploit the phases and their ordering in successful networks to probability distributions which are sampled from to create new solutions.

3.8 Aging Evolution

[2] produces a GA variant in which the population is selected against based upon the "age" of the network. This method was able to set a new SOTA when scaled up on ImageNET. This method maintains a population of networks from which a sub-population of size S are selected and evaluated. The highest scoring network is mutated to create a new child network which replaces the oldest model in the population.

4 Many-Armed Bandit and Early-Stopping

The problem of hyper-parameter optimisation can also be framed as a many armed bandit problem. This approach seeks to find an effective balance between exploration of new hyper-parameter space and exploitation of known, high performing settings.

Early-Stopping is based on the assumption that a partially trained models performance will be predictive of its final performance.[32][47] This assumption can, however, become strained in cases where the variance in convergence time of a set of models is very large. One of the limitation of this method is the final performance with a larger computational budget. This has lead to the incorporation of bandit based strategies into Bayesian and Evolutionary techniques to improve the anytime performance as well as allowing for greater exploration. [32][47][39]

4.1 HyperBand

Hyperband [18] is an popular example of this strategy and has shown the effectiveness of this method, in particular with respect to anytime performance. Hyperband was able to achieve a reduction in training time of an order of magnitude over Bayesian methods while maintaining only a minor reduction in performance.

Hyperband can be considered a resource allocation algorithm for Successive-Halving [10], attacking the problem of balancing the resources allocated to the training of each model (n) with the total number of models trained for a fixed resource budget. Due to the variance of model convergence time between different algorithms or problem domains, this cannot be a static, fit all value. Hyperband effectively performs a grid search of values for n, running a sequence of tests for different values of n within a bounded range.

4.2 ASHA

Asynchronous Successive Halving Algorithm (ASHA) [28] is a parallelisation technique for the Successive-Halving algorithm. This method has been shown to perform as well as or better than other techniques that incorporate early stopping such as PBT, BOHB and Hyperband on a variety of hyper-parameter optimisation and NAS benchmarks.

5 Bayesian Optimisation

Bayesian Optimisation (BO) methods have become popular over the last couple of years due to the SOTA performance they can produce [17] [6]. One of the core weaknesses of BO is the computationally expensive nature of the many BO methods in particular Sequential Model-based Bayesian Optimisation (SMBO).

The objective of BO, in the context of hyper-parameter optimisation, can be described as trying to minimize $x^* = argmin\ f(x)$ where $x \in X$ and $X \subseteq \mathbf{R}^k$, X is bounded and compact and k is the number of hyper-parameters. This method assumes a correlation between observations and endeavours to produce useful points for evaluating, x, by exploiting a model constructed based upon a set of function observations $D = \{(x_n, y_n)\}_{n=1}^{N}$ where $y_n \sim \mathcal{N}(f(x), \sigma^2)$

- assumes correlation between function evaluations
- uses fantasises in [17]
-
-

5.1 Acquisition Function

The acquisition function is used in BO to select the next point in hyper-parameter space at which to sample. This is the component of the system that manages the exploit/explore problem. One of the key problems faced by BO is the sequential nature of the optimisation process. The acquisition function computes a point x in hyper-parameter space as the next point to be queried using data from the posterior model. Using common acquisition functions such as Expected Improvement (EI), Probability of Improvement or Upper Confidence Bound in combination with a Gaussian Process (GP), any subsequent polling

for query points without updating the posterior model will simply return the same point x. There are asynchronous implementations used as a solution to this problem, which commonly involve updating the posterior model as a worker completes an evaluation, producing a new query point from the updated posterior model and re-dispatching the worker. [26][15][6]

5.1.1 Expected Improvement / EI-MCMC

The most common of acquisition function is expected improvement (EI) due to the fact it is considered robust and does not require hyper-parameter tuning of its own. [8] EI-MCMC is an extension of this function which allows for asynchronous parrallelisation based on Monte Carlo estimates of the acquisition function.

5.1.2 Local Penalisation

One intuitive solution to this problem is the application of local penalization to the acquisition function in order to create batches of query points as shown in [9]. However, this method can be consider as "doubly greedy" [11] in cases where a greedy acquisition function is used as the basis for local penalisation.

5.1.3 Thompson Sampling

Thompson Sampling (TS) is one method that has been applied to the problem of BO parrallelisation[26][15]. In TS the next point to be queried (x*) is selected based on a random sampling of the posterior probability distribution. This gives theoretically grounded method for maintaining diversity in synchronise selection of evaluation points without updating the posterior distribution. This was able to outperform other parallel BO methods in hyper-parameter optimisation on the CIFAR-10 data-set. [26] builds on the results of [15] introduction a variant referred to as AEGiS which has a chance of performing a purely explorative evaluation rather than utilising TS.

[11] puts forward a non greedy method for the parallel acquisition of query points which was able to outperform other (greedy) parallel BO systems on synthetic and real world bench-marking functions.

5.2 Posterior Model

The model in a BO problem makes estimations of the modelled function, in our case validation score across hyper-parameter space, while maintaining a measurement of uncertainty. The model returns a mean and variance prediction for a function f(x) at point x. [5]

uncertainty

5.2.1 Gaussian Process

The most common model used in hyper-parameter optimisation is a Gaussian Process (GP)[8]. The GP is a priors over function which produces a distribution for the value of f(x) from a point x. The Gaussian Processes is widely used due to its flexibility as well as the simplicity of computing of many common acquisition functions from it.

5.2.2 Tree-Structured Parzen Estimator

[6] introduces a model based on Parzen Window Density or Kernel Density estimation, known as Tree-Structured Parzen Estimator (TPE). This was able to outperform both random and GP based Bayesian hyper-parameter optimisation on the MNIST data-set. [17] has more recently used this model to great success, leading to its use as the underlying system in the popular Auto-ML tool "HyBandSter" [28]

5.2.3 Bayesian Neural Networks

Another method for modelling the distribution over a function is with the use of a neural network. DNGO [12] is an implementation of this that uses a Deep Neural Network in combination with a Bayesian linear regressor to create an adaptive basis regression as the surrogate function model. This approach has the advantage of scaling linearly, in terms of computation, with the number of observations rather than cubically as with a GP. This method was able to achieve parity with SOTA in hyper-parameter optimisation on CIFAR-10. BOHAMIANN [13] was able to out perform DNGO on a number of hyper-parameter optimisation tasks while supporting native parallelisation, however, due to the complexity of this method, there are a number of key hyper-parameters that need to be effectively turned to produce optimal results.

6 Reinforcement Learning

Short description of reinforcement learning Pros: Strong performance with high enough computational budget Incorporates meta-learning, moving away from a black-box perspective Cons: Very High computational cost

1. [50] Brought RL-NAS to the mainstream producing neural architectures that outperformed human designed SOTA on cifar-10, this required 800 gpu days or something idk

6.1 One-shot Architecture Search

2. ENAS [52] was an expansion upon RL-NAS which was able to improve efficiency by over 1000x and achieve slightly improved results. This was done by training a single acyclic graph within which the entire search-space was

contained. 3. [51] was able to show the importance of dropout in the one-shot method described in [52]

7 Bi-Level/Gradient Descent

[11] STNs - gradient descent for hyper-parameter optimisation
[53] Darts - gradient descent for cell architecture search

8 Meta-Learning

Meta-learning has gained traction as a possible path away from the black-box perspective of hyper-parameter optimisation. This involves the application of machine learning to configuration selection. This has been implemented in various forms with success, such as a type of transfer learning in network embedding applications [24]. This has also been implemented as an interface which allows experts to effectively warm start an auto-ML pipeline with knowledge of effective configurations for similarly structured problems[29]. Another approach has been to train an agent via reinforcement learning to select configurations[28]. These applications of meta-learning are effective at reducing the convergence time however, rarely achieve significantly superior final performance than the methods they are supplementing.[27][29]

9 Research Opportunities

one-shot/hypernetwork applied to GAs or BO
heading=bibintoc, title=Whole bibliography
]