

Farthest Greedy Path Sampling for Two-shot Recommender Search

Yufan Cao

caoyf20@mails.tsinghua.edu.cn
Tsinghua University
Haidian Qu, Beijing Shi, China

Feng Yan

fyan5@uh.com
University of Houston
Houston, Texas, USA

Tunhou Zhang

tunhou.zhang@duke.edu
Duke University
Durham, North Carolina, USA

Wei Wen

wewen@meta.com
Meta AI
Menlo Park, California, USA

Hai Li

hai.li@duke.edu
Duke University
Durham, North Carolina, USA

Yiran Chen

yiran.chen@duke.edu
Duke University
Durham, North Carolina, USA

ABSTRACT

Weight-sharing Neural Architecture Search (WS-NAS) provides an efficient mechanism for developing end-to-end deep recommender models. However, in complex search spaces, distinguishing between superior and inferior architectures (or *paths*) is challenging. This challenge is compounded by the limited coverage of the supernet and the co-adaptation of subnet weights, which restricts the exploration and exploitation capabilities inherent to weight-sharing mechanisms. To address these challenges, we introduce Farthest Greedy Path Sampling (FGPS), a new path sampling strategy that balances path quality and diversity. FGPS enhances path diversity to facilitate more comprehensive supernet exploration, while emphasizing path quality to ensure the effective identification and utilization of promising architectures. By incorporating FGPS into a Two-shot NAS (TS-NAS) framework, we derive high-performance architectures. Evaluations on three Click-Through Rate (CTR) prediction benchmarks demonstrate that our approach consistently achieves superior results, outperforming both manually designed and most NAS-based models.

CCS CONCEPTS

- Information systems → Recommender systems; • Computing methodologies → Neural networks.

KEYWORDS

Recommender System, Neural Architecture Search, Weight Sharing, Path Sampling, Optimization Gap

1 INTRODUCTION

Recommender systems, essential in domains such as entertainment [1], social media [20], and e-commerce [31], have continually evolved to refine user experiences by predicting preferences more accurately. Historical lines of approaches to improve precision include the utilization of collaborative filtering [20, 33], content-based filtering [26, 28], and knowledge-based systems [3, 38]. These methodologies have been pivotal in tailoring recommendations by analyzing and modeling distinct user features, thus facilitating more precise estimations of user behavior. In the evolution of recommender systems, the integration of Deep Neural Networks (DNNs) results in a notable paradigm shift. Initial DNN strategies [11, 22, 27, 37] heavily rely on manual, hand-crafted efforts. The emergence of Neural Architecture Search (NAS) [21, 36, 46]

enables a more systematic and efficient exploration of the DNN design space, reducing the reliance on manual intervention.

Weight-sharing NAS (WS-NAS) [4, 45] achieves remarkable architectural flexibility among various NAS methodologies. WS-NAS substantially reduces computational costs by employing a single *supernet* to represent the entire search space. A supernet is a large, all-encompassing model, from which we zero out certain components or connections to sample a *subnet* that corresponds to a *path* from input to output. The process of sampling a subnet, i.e., a path from the supernet is therefore termed “path sampling”. Weights of the supernet can be naively copied to expedite the evaluation of sampled subnets, based on which NAS identifies presumably top-performing architectures. WS-NAS successfully achieves remarkable architectural flexibility while substantially reducing computational costs.

This methodology opens up opportunities for NAS with highly complex search space settings and prohibitive computational costs. However, it suffers from the “co-adaptation” problem arising from the gap between weights of the supernet and the optimal weights of a subnet yielding its real performance [43]. With such an optimization gap, the incorrectly evaluated top-performing architectures may be unintentionally eliminated during the search, resulting in worse performance of top-performing candidates. We attribute part of this problem to relatively naive existing path sampling strategies [6, 12] where few consider coverage of top-performing paths.

In this paper, we propose “Two-shot Deep Recommender Search” (TS-NAS), a novel NAS approach that advances the path sampling strategy to improve the ranking correlation of different subnets. TS-NAS performs a search on the WS-NAS supernet twice to obtain the optimal subnets. In the first shot, we follow the practice of NAS-Rec [46] to train a supernet and identify a set of well-performing paths, from which we sample a path subset with a maximized diversity measured by Shannon Entropy. The sampled top-performing paths have the “farthest” pairwise distances and “greediest” performance, hopefully having good coverage of the supernet for their diversity, thereby mitigating operator imbalance and weight co-adaptation during search. In the second shot, we train a supernet on Farthest Greedy Paths, after which we perform an evolutionary search for top-performing architectures. We name this path sampling strategy “Farthest Greedy Path Sampling” (FGPS, Figure 1) and highlight it as the major contribution of this work.

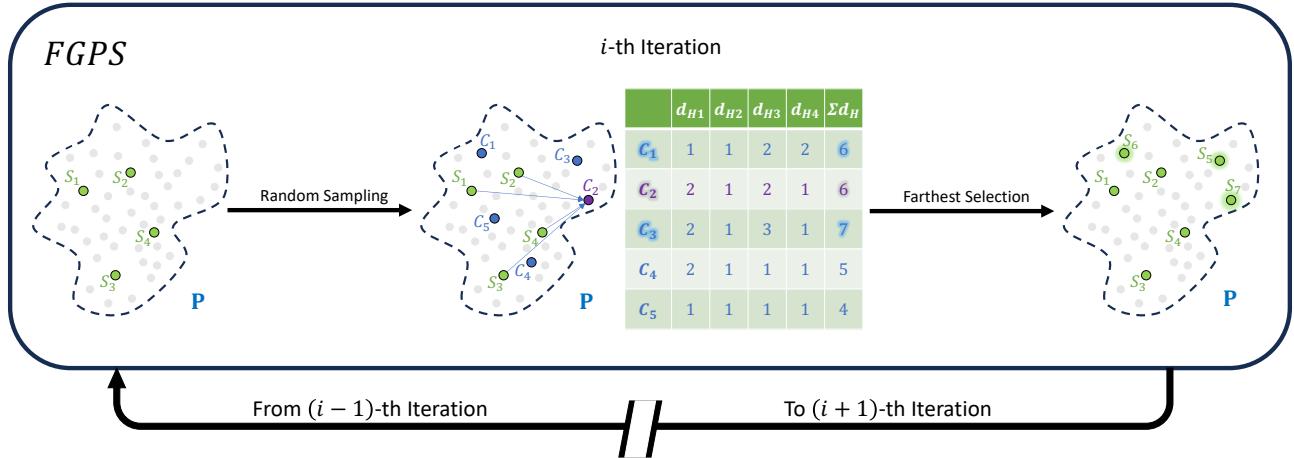


Figure 1: Illustration of Farthest Greedy Path Sampling (FGPS). With the selected path set (green dots) after the $(i-1)$ -th iteration, FGPS samples paths (blue dots) from a path set P (gray dots) and pick out the farthest ones with largest total distance from the selected paths and add them to the selected set as the result of the i -th iteration.

We build our approach on the broad but challenging NASRec search space with heterogeneous building operators, dense connectivity, and elastic dimensions. We thoroughly examine our method on three Click-Through Rates (CTR) prediction benchmarks Criteo Kaggle, Avazu, and KDD Cup 2012. Results demonstrate that our method outperforms all hand-crafted models, and shows an improvement of ~ 0.001 in AUC compared to NAS-crafted models in Criteo and KDD benchmarks. We summarize our contribution as follows:

- We design a heuristic path sampling approach named “FGPS”, which yields a set of paths boasting both diversity and coverage of top-performing architectures in the search space. This set of paths empirically enables the second shot of NAS to find better top-performing architectures.
- We propose a novel NAS approach for recommender systems named “TS-NAS”. It enables training the supernet with paths sampled by FGPS and more accurate evaluation of top-performing recommender system architectures.
- The results show that the architectures our method finds can achieve state-of-the-art (SOTA) performance in most main CTR prediction tasks, verifying its ability to automatically explore and exploit an intricate design space with high heterogeneity.

2 RELATED WORKS

Deep Learning-based Recommender Systems. Deep Neural Networks (DNNs) serve as the backbone models that support end-to-end inference in recommender systems, such as Click-Through Rate (CTR) prediction applications [7, 14, 34]. Starting with Wide & Deep learning [5], the integration of dense and sparse features has played a crucial role in modeling user preferences and optimizing click-through rates. Techniques like Factorization Machines [11], Attention [37], crossing layers [39, 40], and instance-guided masks [41] have been developed to enhance explicit and

implicit feature interactions, thereby establishing strong design priors to drive DNN designs. Another line of research focuses on optimizing the set of feature interactions in recommender systems, including attention/field-aware factorization machines [17, 42], feature interaction search [9, 23]. More recent research [36, 46] utilize Neural Architecture Search (NAS) to expand the scope of exploration in deep recommenders, such as dense connectivity, heterogeneity, and multimodality. Yet, studies on the quality of NAS, such as ranking quality, are rarely conducted in existing works.

Path Sampling in One-shot Search. One-shot NAS [4, 45] utilizes a supernet to represent the search space and carry path sampling to select different set of subnets during training. In one-shot NAS, the quality of path sampling decides the ability of a NAS algorithm to distinguish between superior and inferior architectures [2]. Single-path sampling [12], sandwich rule [45], progressive shrinking [4], and greedy path sampling [16, 44] are past approaches on computer vision problems. Yet, existing approaches do not consider supernet coverage during path selection. This is because the search space within computer vision is relatively smaller and simpler, containing only a single type of block (e.g., MobileNet-v2).

3 METHODOLOGY

We base our search space on the settings of NASRec [46], which is a composition of building operators, connectivity, and elastic dimension.

3.1 Supernet Configuration and its Sampled Path Encoding

Following the definition in [46], a supernet \mathcal{S} with choice blocks from 1 to N is a tuple of connections $C = [C^{(i)}]_{i=1}^N$, operators $O = [O^{(i)}]_{i=1}^N$, and dimensions $D = [D^{(i)}]_{i=1}^N$ over N choice blocks. $C^{(i)}$ defines the connectivity between choice block j and

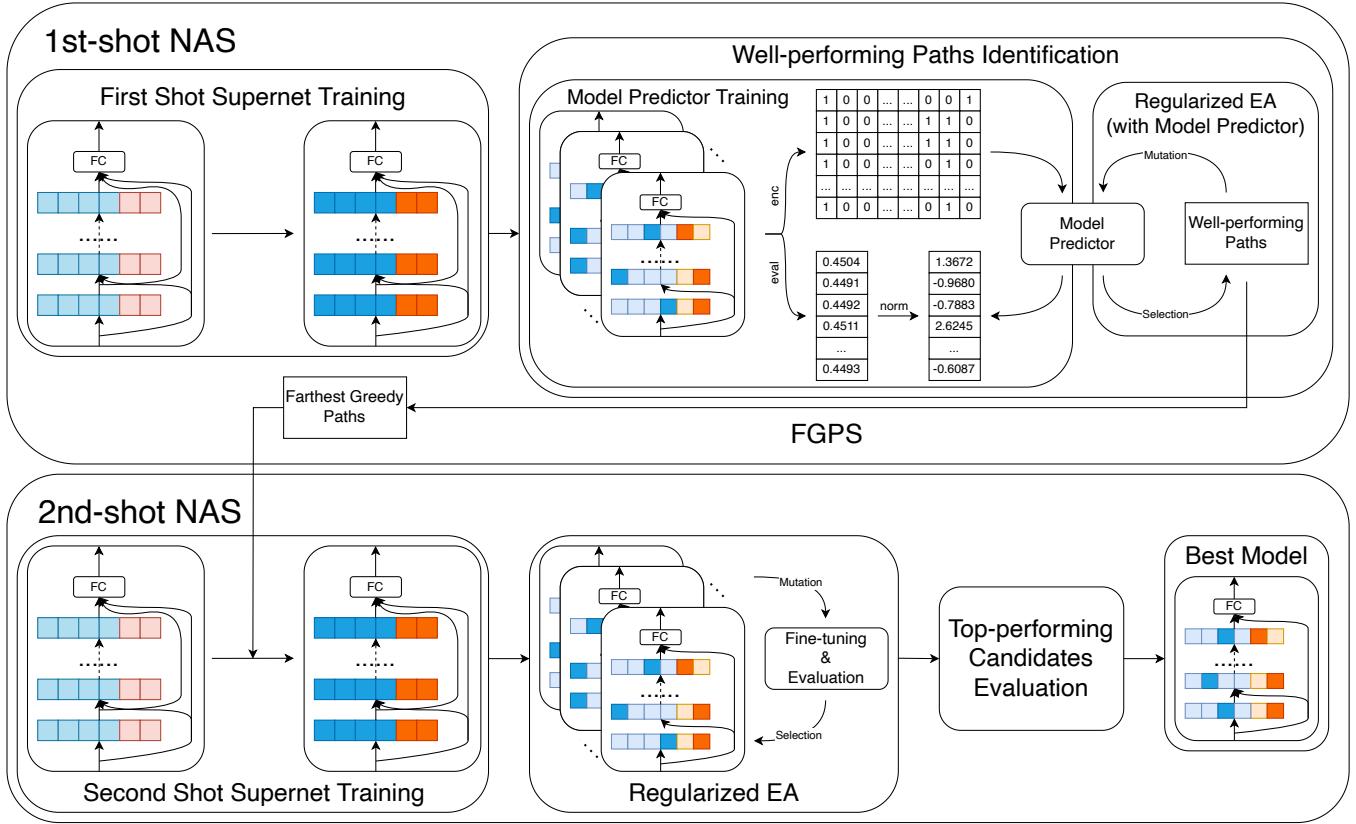


Figure 2: Overview of Two-Shot Neural Architecture Search with Farthest Greedy Path Sampling(FGPS). FGPS prepares paths for training the supernet in the second shot, mitigating the co-adaptation problem and leading to a better best model.

choice block i for all $j < i$ (0 as input), avoiding loops. For example, if there is a connection between choice block 1 and choice block 3 , the outputs of choice block 1 should be taken as part of the inputs of choice block 3 . $O^{(i)}$ defines the set of building operators in choice block i . And $D^{(i)}$ defines the dimension settings in choice block i .

A valid path (i.e., a subnet) $\mathcal{P} = (C_p, O_p, \mathcal{D}_p)$ from input to output within the supernet is a tuple of sampled sets of connections, operators, and dimensions from the search space for each choice block. We formulate the encoding of a sampled path as follows:

- On the inter-block level, we use zero-one encoding to represent the connectivity among blocks. In block j , operator o within a search space of N blocks, the encoding $\epsilon(\mathcal{P}, j, o) \in \{0, 1\}^N$, and $\epsilon[i] = 1_{<i,j>\in C_p^{(j)}}$. For example, the fourth choice block out of seven that takes in raw input, the outputs of the first and the third block as its inputs correspond to an encoding of “110100”. The connectivity is encoded sequentially for all modules and then concatenated to form the inter-block-level encoding.
- On the intra-block level, we pre-assign options for operators, and structural parameters (e.g. #dimensions, with or without dense/sparse interaction), and encode in-block choices with their indices. Encoding is sequentially done for all blocks, concatenated and appended to inter-block-level encoding. We discuss intra-block details in Section 4.1.

3.2 Measurement of Distance and Entropy among Paths

We formulate *distance*(how different one path is from another), and *entropy*(how diverse a path set is) for navigation towards better exploration and exploitation within our encoding space. The path encoding above constructs a bijection between the architecture of a subnet in the search space, and a vectorized representation in the 280-dimensional encoding space. We measure *distance* between paths \mathcal{P}_m and \mathcal{P}_n with Hamming distance [13] Dist_H :

$$\text{Dist}_H(\mathcal{P}_m, \mathcal{P}_n) = \#\text{different_symbols}(\text{enc}(\mathcal{P}_m), \text{enc}(\mathcal{P}_n)). \quad (1)$$

Dist_H evaluates how far path \mathcal{P}_m is from \mathcal{P}_n , allowing the selection of more diverse paths during sampling. Hamming distance is employed with the assumption that different operators hold equivalent chances of achieving high final performances.

Similarly, to measure the diversity of a set of paths $\mathbf{P} = \{\mathcal{P}_m\}_{m=1}^k$, we exploit Shannon entropy [35]. The Shannon entropy Ent_S of \mathbf{P} is formulated as follows:

$$\text{Ent}_S(\mathbf{P}) = \sum_{\text{pos}=1}^{\text{len}(\text{enc})} \left[\sum_{e=1}^{\#\text{enc}(\mathbf{P})[\text{pos}]} -f_{\text{pos}}(e) \log f_{\text{pos}}(e) \right], \quad (2)$$

where $\#enc(\mathbf{P})[\text{pos}]$ denotes the number of different symbols on the pos-th dimension, and

$$f_{\text{pos}}(e) = \sum_{m=1}^k \mathbf{1}(\mathcal{P}_m[\text{pos}] = e)/k \quad (3)$$

is the frequency of symbol e on the pos-th dimension of the encoding. We propose:

THEOREM 3.1. *Given a path set $\mathbf{P} = \{\mathcal{P}_m\}_{m=1}^k$ and an encoding function enc , a maximum Shannon entropy on a path subset $\mathbf{P}^* = \{\mathcal{P}'_n\}_{n=1}^r \subseteq \mathbf{P}$ can be approximated if*

$$\mathbf{P}^* = \arg \max_{\mathbf{P}'} \left[\sum_{i=1}^r \sum_{j=i+1}^r \text{Dist}_H(P'_i, P'_j) \right]. \quad (4)$$

The theorem above indicates that maximizing the pairwise Hamming distance of sampled paths improves the Shannon entropy of the sampled subset, making them more diverse.

3.3 Farthest Greedy Path Sampling

Algorithm 1 Farthest Greedy Path Sampling

Require: A path set \mathbf{P} , an encoding method enc , #budget B per iteration, #paths K added per iteration.
 Selected path set $\mathbf{P}_S \leftarrow B$ different random paths sampled from \mathbf{P}
 Candidate path set $\mathbf{P}_C \leftarrow \emptyset$
while One iteration takes reasonable time **do**
 while $|\mathbf{P}_C| < B$ **do**
 Randomly sample a path \mathcal{P} from \mathbf{P}
 if $\mathcal{P} \notin \mathbf{P}_S \cup \mathbf{P}_C$ **then**
 Add \mathcal{P} to \mathbf{P}_C
 end if
 end while
for \mathcal{P}_i in \mathbf{P}_C **do**
 Compute the sum of Hamming distance between \mathcal{P}_i and all the paths in \mathbf{P}_S
end for
 Pick out K paths from \mathbf{P}_C with highest distance summation and add them to \mathbf{P}_S
 $\mathbf{P}_C \leftarrow \emptyset$
end while
return \mathbf{P}_S

Previous works of weight-sharing one-shot NAS [4, 45] suffer from the so-called “co-adaptation” problem, which arises directly from the design of the method [43]. In each step of training the supernet, they first sample a path randomly from the supernet and perform a one-step gradient descent on the related parameters in the supernet. In the downstream architecture search session, the parameters of related parts of the supernet is directly copied to give a rough evaluation of the performance of a subnet. Though this method efficiently cuts down computational burdens, in evaluation there is an inevitable gap between the supernet parameters and the optimal parameters which yields the real performance of a subnet.

It is impossible to address the co-adaptation problem as the once-trained parameters in the supernet can never be optimal for all sampled paths, so we aim to reduce its impact only on top-performing subnets, with the insight that NAS pays little attention to poor-performing architectures.

We introduce “farthest greedy path sampling” (FGPS) as a foundation to practice the above insight. We confine the path set \mathbf{P} in Theorem 3.1 as a well-performing path set, and aim to find a path subset \mathbf{P}^* with high diversity quantified by Shannon entropy. In FGPS, we use pairwise path distance as a proxy to maximize the Shannon entropy and iteratively sample new paths that are most different from the ones already sampled. The detailed sampling strategy is described in Algorithm 1. As the bridge between first-shot and second-shot NAS, FGPS explores the well-performing path set to gain diversity, and exploits sampled paths for a better model search. Paths sampled by FGPS are referred to as Farthest Greedy Paths.

3.4 Weight-sharing Two-shot NAS Workflow

Equipped with FGPS, we propose a novel approach of NAS, “Weight-sharing Two-shot NAS”(TS-NAS). The TS-NAS workflow can be summarized as below:

- (1) Train first-shot supernet with randomly sampled paths;
- (2) Use an efficient way (see Section 3.5) to roughly identify a set of well-performing paths;
- (3) Perform FGPS on the identified path set to find a diverse well-performing path subset to acquire Farthest Greedy Paths;
- (4) Train second-shot supernet with Farthest Greedy Paths;
- (5) Perform evolutionary search on sampled paths from trained second-shot supernet to generate top-performing path candidates;
- (6) Evaluate top-performing candidates and pick out the best architecture.

With supernet training on a well-performing diverse path set, parameter updating is directed towards the optimal parameter of top-performing paths, narrowing the aforementioned parameter gap. Therefore, it mitigates the “co-adaptation” problem of the top-performing paths.

3.5 Identification of a Well-performing Path Set

With a search space containing $\sim 10^{33}$ different architectures, it is impossible to identify a large set of well-performing paths with direct evaluation of the subnets. To address this challenge, we adopt the encoding space.

Given a 1-shot trained supernet, we randomly sample a lot of paths \mathbf{P}_{rand} from it, fine-tune them (which is computationally affordable), and get an evaluation $\text{logloss}(\mathbf{P}_{\text{rand}})$. Then we perform enc to map these sampled paths into the encoding space, generating $\text{enc} \mapsto \text{logloss}$ mappings inside \mathbf{P}_{rand} . We then train a simple MLP with embedding to learn the mappings. As the encoding function has constructed a mapping from paths to encodings, such method constructs a direct mapping from model architectures to their predicted performance and effectively lowers the computational budgets.

With this mapping from paths to predicted performances as a proxy, we can now identify a large set of well-performing paths with

ease. We use an evolutionary algorithm to iteratively generate new paths and keep the top ones with the highest predicted performance.

4 EXPERIMENTS

We first delineate our basic experimental settings including search space and dataset configuration, hyperparameters and baseline implementations. Then we provide a detailed step-by-step Two-shot NAS Workflow in very fine grain, followed by a comparative analysis of experimental results with baselines. Finally we investigate how FGPS improves Shannon Entropy during path sampling, and the role of Shannon Entropy throughout the workflow.

4.1 Experimental Setup

4.1.1 Search Space and Datasets. Our method is built upon the open-source implementation provided by NASRec[46], from which we adopt the defined search space.

At a macroscopic level, the search space consists of seven choice blocks, permitting arbitrary connections from preceding blocks to subsequent ones. At a microscopic level, the functionality within each block is governed by specific operators. These operators can be categorized based on their input and output types:

- Operators handling dense inputs to produce dense outputs: This category includes the Fully-Connected (FC) layer, Sigmoid Gating (SG) layer, and Sum layer.
- Operators processing sparse inputs to yield sparse outputs: This category includes the Embedded Fully-Connected (EFC) layer and the Attention (Att) layer.
- Operators managing a combination of dense and sparse inputs to generate dense outputs: This category includes the Dot-Product (DP) layer.

In scenarios where multiple operators are concurrently active within a dense or sparse branch, their outputs are aggregated via summation. In our implementation, the DP layer has been modified with an expansion of the dense dimension by a factor of 16.

Within each choice block, mergers exist to integrate "dense" and "sparse" data. Specifically, a "dense-to-sparse" merger accepts dense outputs, processes them via a Fully Connected (FC) layer, and then reshapes the outcome into a 3D sparse tensor. This tensor is subsequently concatenated to the sparse outputs. Conversely, a "sparse-to-dense" merger takes as input sparse outputs, processes them using a Factorization Machine (FM) [30], and then adds the resulting output to the dense outputs.

We perform a Neural Architecture Search and model evaluation on three click-through-rate (CTR) prediction benchmarks: Criteo Kaggle¹, Avazu², and KDD Cup 2012³. We used a split of 80% training data, 10% validation data, and 10% data for each benchmark. We use Stratified K-fold [29] to obtain the dataset splits with a random seed of 2018, ensuring the consistency with existing works such as AutoInt [37], AutoCTR [36], and NASRec [46].

All model performances are assessed using Logarithmic Loss as the evaluation metric.

¹<https://www.kaggle.com/c/criteo-display-ad-challenge>

²<https://www.kaggle.com/c/avazu-ctr-prediction>

³<https://www.kaggle.com/c/kddcup2012-track2>

4.1.2 Training Settings. To ensure a fair comparison and better demonstrate our efforts in architecture search, we employ a consistent training protocol for all architectures, deliberately avoiding hyperparameter tuning. Specifically, we employed the Adagrad optimizer [8] with an initial learning rate of 0.12. A batch size of 256 was utilized for the Avazu and Criteo datasets, while a size of 512 was designated for the KDD dataset. All models were optimized over one single epoch, and a cosine learning rate schedule [24] is added to gradually decay the learning rate to 0 at the end of the training epoch.

4.1.3 Baselines. We implement all baselines based on open-source code and paper descriptions. Most models have 2 layers in dense processing layers and 6 layers in overarching layers, as recommended in AutoCTR [36]. We carefully tune the architecture of each baseline to ensure a competitive implementation over all CTR benchmarks.

- **DLRM** [27] has a two-tower architecture composed of 2 MLP layers in dense feature processing and a 6-layer MLP in high-order feature processing. All hidden layers have 1024 units.
- **DeepFM** [11] has a wide & deep architecture that integrates the benefits of FM for factorization and deep learning for abstract feature learning. The dense feature processing is done with a two-layer MLP, and a 6-layer MLP is for high-order feature extraction. All hidden layers have 1024 units.
- **xDeepFM** [22] is a two-tower model featuring a Compressed Interaction Network (CIN) to perform vector-wise explicit feature generation. Besides CIN, xDeepFM has 2 MLPs in dense feature processing and 6 MLPs in high-order feature processing. All hidden layers have 1024 units.
- **AutoInt+** [37] extends the AutoInt architecture by incorporating additional techniques such as feature selection and self-attention mechanisms to enhance model performance, particularly in handling high-dimensional feature spaces. All hidden layers have 1024 units.
- **SerMaskNet** [41] employs a series of masking techniques within a deep neural network structure and across MaskBlocks to enable more efficient and effective feature interaction and selection, fostering better learning and generalization. SerMaskNet has three MaskBlock in series with all hidden layers having 1024 units.
- **ParaMaskNet** [41] operates similarly to SerMaskNet but introduces parallel masking networks to capture different kinds of features for a nuanced pattern capturing. ParaMaskNet has three MaskBlock in parallel with all hidden layers having 1024 units.
- **DCN-v2** [40] stands for Deep and Cross Network version 2, an evolution of the original DCN, featuring enhanced deep and cross combination for more powerful feature interaction learning and representation. All hidden layers have 1024 units.
- **NASRecNet** [46] leverages weight-sharing NAS within the realm of recommender systems and employs "Single-operator Any-connection" sampling, evolutionary algorithms to approximate the best model. We follow the open-source implementation for this baseline.

Table 1: Performance of TS-NAS with FGPS on Three Click-Through Rate Prediction tasks.

	Method	Criteo		Avazu		KDD	
		logloss	AUC	logloss	AUC	logloss	AUC
Manually-designed Models	DLRM	0.4408	0.8105	0.3760	0.7864	0.1528	0.7980
	DeepFM	0.4402	0.8111	0.3798	0.7797	0.1541	0.7905
	xDeepFM	0.4400	0.8114	0.3777	0.7836	0.1544	0.7886
	AutoInt+	0.4447	0.8062	0.3796	0.7803	0.1499	0.8120
	SerMaskNet	0.4408	0.8104	0.3782	0.7828	0.1496	0.8128
	ParaMaskNet	0.4402	0.8111	0.3772	0.7843	0.1496	0.8130
	DCN-v2	<u>0.4389</u>	<u>0.8125</u>	0.3789	0.7814	0.1548	0.7864
NAS-crafted Models	NASRec@Small	0.4391	0.8123	<u>0.3751</u>	<u>0.7880</u>	0.1489	0.8165
	NASRec@Full	0.4394	0.8119	0.3749	0.7883	<u>0.1487</u>	<u>0.8168</u>
	TS-NAS w/FGPS@Full	0.4387	0.8127	0.3755	0.7874	0.1485	0.8179

4.2 Two-Shot NAS Workflow

We use the Criteo dataset as a representative example and explain our workflow through each phase.

4.2.1 First-shot Supernet Training. During the initial phase of the first-shot NAS, the supernet undergoes training using paths sampled at random. For the first 15,000 steps, there is a progressively diminishing probability that the entire supernet is trained to avoid model collapse. We adopt the “Single-operator Any-connection” strategy [46] for this sampling approach. In detail, each choice block, we select one dense operator and one sparse operator, and we permit connections between all blocks. This strategy is specifically tailored for the subsequent model predictor training phase for its superior performance that the ranking of model predictions closely aligns with their actual value rankings. For the training process, we employ the Adam optimizer [19] with an initial learning rate of 0.001 and a batch size of 512.

4.2.2 Model Predictor Training. After training the supernet, we pick paths from it at random to create a set denoted as P_R . To identify a large set of well-performing paths efficiently, we train a model predictor using this set.

First, we apply the encoding function enc on the path set and acquire $\text{enc}(P_R)$ as a matrix containing the tokenized path configurations. Then, we prepare the models for these paths using the weights from the trained supernet and fine-tune them by training for another 0.5K steps. This gives us a dataset that shows how different encoded paths perform in terms of Log Losses.

For each path, we embed each token in the encoding into 15-dimensional vectors with a look-up table. After combining these vectors, we process them using a Multi-Layer Perceptron [32]. The goal is to adjust the model predictor so that its predictions match the normalized Log Losses of the paths. We choose the best hyperparameter settings for the model based on the test Kendall’s τ coefficient [18]. Finally, we train the model using the entire dataset to prepare it for the next steps.

4.2.3 Identification of a Well-performing Path Set. Utilizing the model predictor as a surrogate, we efficiently identify a vast collection of high-performing paths using a regularized Evolutionary Algorithm (EA) [10, 15]. We begin by initializing a population of 20K paths. In each iteration, 51 paths are randomly sampled from

this population. The top-performing path from this sample is then selected and mutations are applied to generate 64 child paths. Based on their predicted normalized log losses, the top 10 child paths are chosen. This process is repeated for 20K generations, leading to a final set of 200K paths. Without the model predictor, obtaining such a large set of well-performing paths would be prohibitively expensive.

4.2.4 Farthest Greedy Path Sampling. From the well-performing path set, we begin by eliminating redundant encodings. Due to the long-tailed distribution of the predicted performance, we also exclude the bottom $\sim 7.5\%$ of the paths. Starting with an initial set of 200 paths selected, we draw 200 paths from those not yet visited during each iteration. From this drawn subset, we retain the 20 paths that have the greatest total distance from all paths in our selected set. This procedure results in $\sim 5.5\text{K}$ distinct paths after iterations, namely Farthest Greedy Paths. The role of Shannon Entropy in this selection process is further explored in section 4.4. At this point, we are poised to commence the second shot of our Neural Architecture Search.

4.2.5 Second-shot Supernet Training. In the second-shot training, we utilize the Farthest Greedy Paths to train the supernet. Throughout the initial 15,000 steps, we maintain a warm-up process for the supernet, with the warm-up probability gradually reducing to zero. To ensure a fair comparison with baseline methods and to emphasize the significance of FGPS, we use the AdaGrad optimizer [8] with an initial learning rate of 0.12, consistent with the settings of the baselines.

4.2.6 Second-shot Evolutionary Algorithm. In the second shot of the regularized evolutionary algorithm, we aim for a more detailed assessment of the subnets. Each subnet undergoes fine-tuning for 0.5K steps and is then evaluated on the test dataset. The process begins by initializing a population comprising 128 subnets. From a sample of 64, the best-performing model is identified, and 8 mutations from this model are subsequently evaluated. This procedure is carried out over 240 generations, after which the outcomes are sorted based on their evaluated losses in the test dataset.

4.2.7 Top-performing Candidates and Best Model Evaluation. We focus on the top 15 subnets identified by their minimal test losses, and further fine-tune and evaluate them. The model exhibiting

the lowest test loss is then selected as the best. Subsequently, we assess this best model starting from scratch, employing a batch size of 256, the AdaGrad optimizer, and an initial learning rate of 0.12 for one epoch. This approach is consistent with the evaluation configurations of the baselines. The resulting losses are presented in Table 1.

4.3 Comparative Analysis

Our approach significantly surpasses the majority of manually designed recommender systems across all datasets. It also consistently outperforms models discovered through NAS, with a stable margin of log loss observed on the Criteo and KDD datasets. Additionally, our method achieves an improvement of approximately 0.001 in the AUC metric when compared to the baselines. This underscores our proposition that FGPS mitigates the optimization gap and co-adaptation problem, offering superior performance to one-shot weight-sharing NAS for recommender systems without the need for parameter tuning. This can also be corroborated in Section 5.1.

An interesting observation is our method’s inability to exceed NASRec’s performance on the Avazu dataset. We postulate that this discrepancy arises from the Avazu dataset’s unique characteristic of lacking dense inputs. This results in limited interactions between dense and sparse components, which are precisely where the search space’s heterogeneity is most demonstrated. Consequently, our search algorithm might become trapped in a local optimum, incorporating redundant dense/sparse interactions. We have scrutinized the top paths identified by the second-shot EA to corroborate this conjecture in Section 5.3.

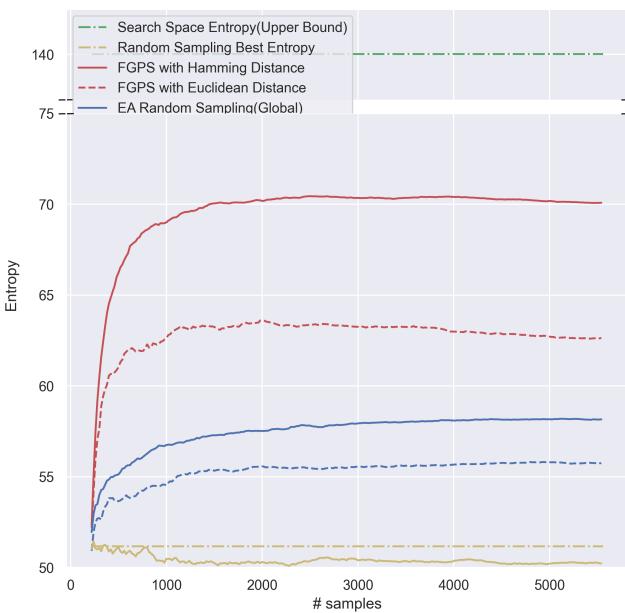


Figure 3: Comparison of Shannon Entropy of path sets iteratively sampled according to different strategies.

4.4 FGPS and Shannon Entropy

We take the NAS on Criteo dataset in the full NASRec [46] search space as an example and investigate Farthest Greedy Path Sampling.

4.4.1 Shannon Entropy Improves with FGPS. To validate the efficacy of FGPS in enhancing path set diversity measured by set Shannon Entropy, we benchmarked it against several path sampling baseline strategies. In the EA random sampling, paths are sampled randomly multiple times, retaining those with the highest local/global entropy of the sample/selected set. Figure 3 portrays a clear distinction in performance between FGPS and its counterparts, with FGPS manifesting superior sampling efficacy. Notably, FGPS utilizing Hamming Distance surpasses its Euclidean Distance counterpart, echoing that a uniform categorical distribution is a more apt surrogate optimum for Shannon Entropy. All heuristic sampling methods significantly outperform random sampling, validating their statistical significance.

Using a sampling strategy as FGPS with Hamming Distance, the Shannon Entropy of the selected path set rapidly increases to approximately 70, and terminates with $\sim 5.5K$ paths sampled, attesting the effectiveness of FGPS. As the sampling progresses, the entropy of the FGPS-sampled path set slightly decreases, suggesting an efficient attainment of optimal Shannon Entropy by FGPS at an early sampling stage.

4.4.2 Shannon Entropy throughout the Workflow. Using the encoding approach described in Section 3.1, we calculate the maximum Shannon Entropy for our search space to be 140.0. This value corresponds to a set of paths that has a uniform distribution for each position in the encoding vector. After applying the first-shot Regularized Evolutionary Algorithm, the resulting path set has an entropy of approximately 50.5. We attribute this decrease in Shannon Entropy to the search space narrowing down to a biased set of paths that perform better. Given this, FGPS’s goal is to elevate the entropy, ensuring a diverse set of paths even within a performance-biased path set. This allows for a broader exploration

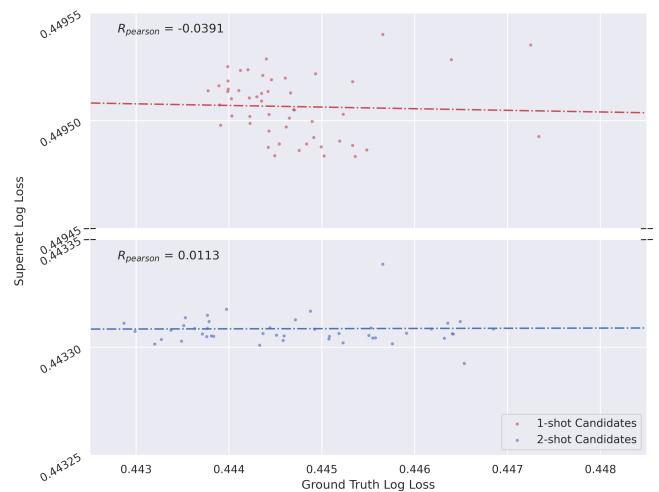


Figure 4: Ranking analysis of top-performing paths found by 1-shot Weight-sharing NAS and TS-NAS w/FGPS.

of the supernet, while focusing on the top-performing subnets for a better exploitation of them in the second shot NAS.

5 DISCUSSION

We provide the ranking analysis, visualization of top-performing paths for distribution analysis and review the top-performing paths.

5.1 Ranking Analysis

We investigate the top 50 subnets found by the evolutionary algorithm by 1-shot weight-sharing NAS and TS-NAS with FGPS to verify that FGPS improves the rankings among the top-performing architectures. We denote Supernet Loss as the test loss after fine-tuning the parameters in the trained supernet, and Ground-Truth Loss as the test loss after training the subnet from scratch. We observe an evident gap between the supernet losses, and an improvement in Pearson correlation in Figure 4. Considering the high sensitivity of the top-performing architectures, this attests that our method mitigates the optimization gap for the top-performing models. Moreover, the ground-truth performance distribution of top candidates discovered by TS-NAS with FGPS exhibits an evident shift towards lower losses compared to one-shot results, proving the effectiveness of FGPS for TS-NAS.

5.2 Top-performing Paths Visualization

We compare the top candidates found by the regularized evolutionary algorithm in the first shot and the second shots. We randomly sample 200 paths, and fit a UMAP [25] together with first-shot and second-shot candidates in Figure 5. We observe a much higher

overlap between random paths and 1-shot candidates, suggesting a higher statistical significance of paths found by TS-NAS. We also observe that candidate paths found by TS-NAS lie on a more continuous manifold, while candidate paths found by 1-shot NAS fall into many discrete clusters. This confirms the ability of TS-NAS with FGPS to exploit what has been explored in the first shot and focus on the most promising paths in the search space.

5.3 Top-performing Model Architecture

Compared with the best model found by 1-shot weight-sharing NAS in Criteo dataset on the same search space, our method finds a model with much more effective modules. Specifically, only one out of the 14 operators is unused in our model, in stark contrast to the six unused operators observed in the best model from the 1-shot weight-sharing NAS.

Our method did not achieve superior performance in the relatively simpler Avazu dataset, which lacks dense inputs and necessitates minimal communication between dense and sparse features. Upon examining the top models we discovered, we observed that the majority employed Sigmoid Gating for over half of the dense operators. This led to an abundance of redundant or ineffective dense feature processing. This limitation can be attributed to the model predictor’s challenge in semantically discerning each operator and a Farthest Greedy Path set that overly prioritizes dense/sparse interactions. That our method roughly made par with NASRec [46] in Avazu suggests that employing FGPS might not be essential in a less heterogeneous search space. Instead, the strength of FGPS comes when exploring more diverse and intricate search spaces.

5.4 FGPS for TS-NAS as a Generic WS-NAS Paradigm

The inherent challenge of coadaptation is deeply rooted in the design of Weight-sharing NAS and is not exclusive to Recommender Systems. We propose FGPS for TS-NAS as a universal framework tailored for weight-sharing NAS, especially when dealing with a search space characterized by high heterogeneity and complexity.

The true power of FGPS lies in its capacity to balance the dual objectives of exploration and exploitation. This dual capability makes FGPS a pivotal component in navigating the intricate balance between exploration and exploitation in NAS search space.

6 CONCLUSION

In this study, we propose a new path sampling strategy “Farthest Greedy Path Sampling”(FGPS), and a two-shot Neural Architecture Search(TS-NAS) workflow tailored for deep recommender systems. TS-NAS incorporates FGPS to adeptly balance path quality and diversity, thus maximizing the effectiveness of the search process. Through rigorous evaluations on multiple Click-Through Rate (CTR) prediction benchmarks, our approach consistently showcased superior performance against all manually designed and most NAS-derived recommender models. This work not only advances the state-of-the-art in NAS for recommenders, introduces a new paradigm of weight-sharing NAS in highly heterogeneous search spaces, but also sets a new benchmark for future research in this domain.

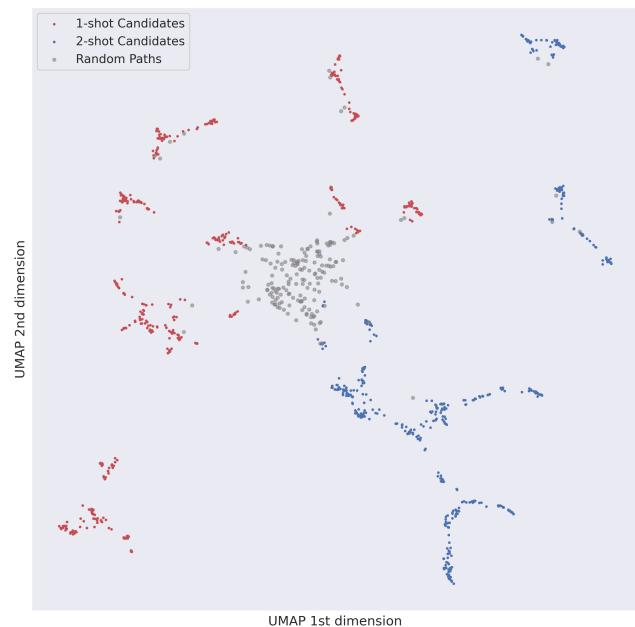


Figure 5: UMAP plot of path encodings of first-shot, second-shot candidates and randomly sampled 200 paths. A lower overlap with random paths suggests a higher statistical significance.

REFERENCES

- [1] Robert M Bell and Yehuda Koren. 2007. Lessons from the Netflix prize challenge. *Acm Sigkdd Explorations Newsletter* 9, 2 (2007), 75–79.
- [2] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. 2018. Understanding and simplifying one-shot architecture search. In *International conference on machine learning*. PMLR, 550–559.
- [3] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12 (2002), 331–370.
- [4] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2019. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791* (2019).
- [5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [6] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. 2021. FairNAS: Rethinking Evaluation Fairness of Weight Sharing Neural Architecture Search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12239–12248.
- [7] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [8] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12, 7 (2011).
- [9] Chen Gao, Yinfeng Li, Quanming Yao, Depeng Jin, and Yong Li. 2021. Progressive feature interaction search for deep sparse network. *Advances in Neural Information Processing Systems* 34 (2021), 392–403.
- [10] David E Goldberg. 2013. *Genetic algorithms*. pearson education India.
- [11] Huirong Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
- [12] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. 2020. Single path one-shot neural architecture search with uniform sampling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI* 16. Springer, 544–560.
- [13] R. W. Hamming. 1950. Error detecting and error correcting codes. *The Bell System Technical Journal* 29, 2 (1950), 147–160. <https://doi.org/10.1002/j.1538-7305.1950.tb00463.x>
- [14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [15] John H Holland. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
- [16] Tao Huang, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. 2022. Greedynasv2: Greedier search with a greedy path filter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11902–11911.
- [17] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM conference on recommender systems*. 43–50.
- [18] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
- [19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [21] Ravi Krishna, Aravind Kalaiar, Bichen Wu, Maxim Naumov, Dheevatsa Mudigere, Misha Smelyanskiy, and Kurt Keutzer. 2021. Differentiable NAS Framework and Application to Ads CTR Prediction. *arXiv preprint arXiv:2110.14812* (2021).
- [22] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.
- [23] Bin Liu, Chenxu Zhu, Guilin Li, Weinan Zhang, Jincai Lai, Ruiming Tang, Xiqiang He, Zhenguo Li, and Yong Yu. 2020. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2636–2645.
- [24] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
- [25] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [stat.ML]*
- [26] Raymond J Mooney and Loriene Roy. 2000. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*. 195–204.
- [27] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. 2019. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091* (2019).
- [28] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*. Springer, 325–341.
- [29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [30] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.
- [31] Francesca Ricci, Lior Rokach, and Bracha Shapira. 2010. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.
- [32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
- [33] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.
- [34] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web*. 111–112.
- [35] C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [36] Qingquan Song, Dehua Cheng, Hanning Zhou, Jiyan Yang, Yuandong Tian, and Xia Hu. 2020. Towards automated neural interaction discovery for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 945–955.
- [37] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1161–1170.
- [38] Karen HL Tso-Sutter, Leandro Balby Marinho, and Lars Schmidt-Thieme. 2008. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 2008 ACM symposium on Applied computing*. 1995–1999.
- [39] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD’17*. 1–7.
- [40] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.
- [41] Zhiqiang Wang, Qingyun She, and Junlin Zhang. 2021. MaskNet: Introducing feature-wise multiplication to CTR ranking models by instance-guided mask. *arXiv preprint arXiv:2102.07619* (2021).
- [42] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional factorization machines: Learning the weight of feature interactions via attention networks. *arXiv preprint arXiv:1708.04617* (2017).
- [43] Lingxi Xie, Xin Chen, Kaifeng Bi, Longhui Wei, Yuhui Xu, Lanfei Wang, Zhengsu Chen, An Xiao, Jianlong Chang, Xiaopeng Zhang, and Qi Tian. 2021. Weight-Sharing Neural Architecture Search: A Battle to Shrink the Optimization Gap. *ACM Comput. Surv.* 54, 9, Article 183 (oct 2021), 37 pages. <https://doi.org/10.1145/3473330>
- [44] Shan You, Tao Huang, Mingmin Yang, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Greedynas: Towards fast one-shot nas with greedy supernet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1999–2008.
- [45] Jiahui Yu, Pengchong Jin, Hanxiao Liu, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Thomas Huang, Xiaodan Song, Ruoming Pang, and Quoc Le. 2020. Bignas: Scaling up neural architecture search with big single-stage models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII* 16. Springer, 702–717.
- [46] Tunhor Zhang, Dehua Cheng, Yuchen He, Zhengxing Chen, Xiaoliang Dai, Liang Xiong, Feng Yan, Hai Li, Yiran Chen, and Wei Wen. 2022. NASRec: Weight Sharing Neural Architecture Search for Recommender Systems. *arXiv preprint arXiv:2207.07187* (2022).