DC-NAS: Divide-and-Conquer Neural Architecture Search for Multi-Modal Classification

Xinyan Liang[1], Pinhan Fu[1], Qian Guo[2], Keyin Zheng[1], Yuhua Qian[1]*

[1] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China
[2] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
{liangxinyan48, fupinhan168, czguoqian, zhengkeyin1221}@163.com, jinchengqyh@126.com

Abstract

Neural architecture search-based multi-modal classification (NAS-MMC) methods can individually obtain the optimal classifier for different multi-modal data sets in an automatic manner. However, most existing NAS-MMC methods are dramatically time consuming due to the requirement for training and evaluating enormous models. In this paper, we propose an efficient evolutionary-based NAS-MMC method called divide-and-conquer neural architecture search (DC-NAS). Specifically, the evolved population is first divided into $k + 1$ sub-populations, and then k sub-populations of them evolve on k small-scale data sets respectively that are obtained by splitting the entire data set using the k-fold stratified sampling technique; the remaining one evolves on the entire data set. To solve the sub-optimal fusion model problem caused by the training strategy of partial data, two kinds of sub-populations that are trained using partial data and entire data exchange the learned knowledge via two special knowledge bases. With the two techniques mentioned above, DC-NAS achieves the training time reduction and classification performance improvement. Experimental results show that DC-NAS achieves the state-of-the-art results in term of classification performance, training efficiency and the number of model parameters than the compared NAS-MMC methods on three popular multi-modal tasks including multi-label movie genre classification, action recognition with RGB and body joints and dynamic hand gesture recognition.

Introduction

In recent years, multi-modal learning has emerged as a powerful approach to enhance the performance of various machine learning tasks by leveraging complementary information from multiple data modalities (Xu et al. 2023c; Wen et al. 2023; Zhuge et al. 2022; Xu et al. 2023a,b; Jiang et al. 2023; Zhang et al. 2022; Han et al. 2023; Zhou et al. 2022). This fusion of information from different sources, such as images, texts, audios, and other forms of data, has shown great potential in tackling complex real-world problems, ranging from image and speech recognition to healthcare and autonomous driving. However, achieving an effective and efficient fusion of multi-modal features remains a challenging task.

In the pursuit of optimal multi-modal feature fusion strategies, researchers have turned their attention to neural architecture search (NAS) that is one state-of-the-art technique for automating the process of designing high-performing neural network architectures. NAS has demonstrated remarkable success in discovering optimal multi-modal feature fusion strategies that outperform hand-crafted ones (Liang et al. 2021). For instance, in specific domains, MMnas (Yu et al. 2020) applies NAS to multi-modal learning with the aim of discovering Transformer model architectures for visual-text alignment, while MMIF (Peng et al. 2020) seeks the optimal CNN structures for extracting multi-modality image features from tomographic scans. MFAS (Perez Rua et al. 2019) and BM-NAS (Yin et al. 2022), on the other hand, are two more general frameworks capable of efficiently searching for multi-modal fusion strategies and enhancing the performance of multi-modal classification tasks.

Although the existing multi-modal NAS methods have achieved promising results in various multi-modal tasks, most of them need to train extensive multi-modal neural networks in each update step, tending to cost more time than non-NAS ones. The gradient-based multi-modal NAS methods greatly improve search efficiency, but their search spaces heavily depend on the super-networks that are predefined. As the number of modalities increases and the scale of data grows, it is necessary to propose multi-modal NAS methods with high computational efficiency and large search space.

In this paper, we propose a population-based multi-modal NAS method called divide-and-conquer neural architecture search (DC-NAS) with a high computational efficiency and large search space. DC-NAS can efficiently adapt to various multi-modal feature fusion strategies and learns DNN networks to tackle diverse multi-modal classification tasks. It primarily relies on evolutionary NAS as the main framework where features extracted from various single-modal DNNs and basic fusion operators are encoded in a tree-based representation. Following the principles of biological evolution, it iteratively searches for the optimal individual to construct the best multi-modal feature fusion network. During the population iteration process, DC-NAS employs a divide-and-conquer search strategy, breaking down the larger problem into a set of smaller and simpler subproblems that are solved iteratively. This strategy has been widely applied to

*The corresponding author.

large-scale optimization problems (Guo, Qian, and Liang 2022; Bi, Xue, and Zhang 2021). With the guidance of the strategy, we first split the training dataset and the population into multiple small-scale datasets and sub-populations, respectively. Then, one sub-populations of them uses the entire training dataset, and the rest ones use one split small-scale dataset for training and evaluation during the evolution process. This intuitively reduces training time since the efficiency limitation of searching for the optimal solution using evolutionary algorithms lies in evaluating each individual. The sub-populations except one in DC-NAS are only trained using different subsets sampled from training data for multi-modal feature fusion learning, which may result in performance loss. In this work, the issue is addressed by exchanging the knowledge between sub-populations via two specially-designed knowledge bases in each generation.

It is worth noting that the idea that employing multiple small populations with information exchange between them for the search, resembles the island-based evolutionary algorithms (Lardeux and Goffon 2010), makes DC-NAS a special case of such methods. However, unlike these approaches, DC-NAS divides the training set into several mutually exclusive subsets and incorporates knowledge transfer, aiming to improve efficiency while maintaining performance to the best possible extent.

The contributions of our work are as follows:

- In order to efficiently utilize diverse multi-modal features for multi-modal classification tasks, we propose a novel multi-modal method DC-NAS, which adaptively selects and fuses features from various modalities to find the optimal fusion network.
- DC-NAS where most individuals evolve with the partial data, only few individuals evolve with the entire data, and knowledge is allowed to exchange between them achieves the comparable performance with one where all individuals evolve with the entire data. This design theoretically and empirically reduces the computation time.
- The extensive comparison experiments on three multi-modal tasks show that DC-NAS achieves competitive performance with reduced search time and fewer model parameters compared to the the state-of-the-art multi-modal feature fusion methods.

Related Work

Neural Architecture Search: NAS is a hot research field that aims to find the optimal neural network architecture through automated methods. In recent years, several approaches related to NAS have emerged and can be broadly categorized into three types: gradient-based methods(Liu, Simonyan, and Yang 2019), reinforcement learning-based methods(Zoph and Le 2017) and evolution-based methods (Liang et al. 2021).

The gradient-based methods require the construction of a super network in advance and the manual design of the search space, deviating from the goal of automation (Yuan et al. 2023). Moreover, compared to evolution-based methods, gradient-based methods have a relatively small solution space and can easily get trapped in local optima (Dong et al. 2021); Reinforcement learning-based methods formulate the structure search problem as a Markov decision process and use reinforcement learning algorithms to learn search policies. In the search process, reinforcement learning algorithms interact with the environment to collect feedback signals and update search policies to obtain better neural structures. Although these methods have achieved satisfactory performance, they often require significant computational resources, making them challenging to apply widely.

Evolutionary-based methods(Yuan et al. 2023) employ evolutionary algorithms to search for neural architectures. They start with an initial population and use evolutionary operations such as selection, crossover, and mutation to iteratively improve and optimize the structures, or adopt particle swarm optimization to search for model architectures. This approach typically guides the evolution process by evaluating the performance of each candidate structure to find the best one. However, a major limitation of evolutionary algorithms is the need to evaluate multiple individuals, which leads to significant training overhead. To overcome this constraint, this paper proposes a novel divide-and-conquer-based method to improve training efficiency without substantially compromising model performance. This method is applied to the multi-modal feature fusion neural architecture search for multi-modal classification tasks.

Multi-Modal Fusion: In the context of deep neural networks, multimodal fusion techniques can generally be categorized into three types: early fusion, late fusion, and hybrid fusion. Early fusion involves combining low-level features, late fusion combines decision-level outputs, while hybrid fusion combines both early and late fusion to achieve enhanced results. So far, various effective combination techniques such as tensor pooling (Hou et al. 2019), Dempster-Shafer theory (Liu et al. 2023) and association-based fusion (Liang et al. 2022, 2023) have been proposed. Deep neural networks, as leading feature extractors, often produce extensive features for each modality data, making manual selection for feature fusion a challenging task. There are two kinds strategies to deal to this issue. The first strategy is to perform fusion at multiple intermediate layers based on predefined fusion rules such as CentralNet (Vielzeuf et al. 2019) and MMTM (Vaezi Joze et al. 2020). While these methods have demonstrated promising performance on multiple tasks, they often lead to an increase in model parameters.

The second strategy is to transfer the original task into a neural architecture search (NAS) task. For example, MFAS (Perez Rua et al. 2019) introduced NAS methods to multi-modal learning for automatically giving a satisfying solution to multi-modal feature fusion. However, the sequential model-based optimization algorithm needs to train and evaluate lots of deep neural networks, leading to searching inefficiency. To address this issue, the gradient-based NAS methods such as DARTS (Liu, Simonyan, and Yang 2019), MMIF(Peng et al. 2020), 3D-CDC(Yu et al. 2021), and BM-NAS (Yin et al. 2022) have been proposed. Such methods train a super-network instead of lots of neural networks, reducing the searching time. However, the structures of super-networks are limited. For example, BM-NAS (Yin et al. 2022) restricts the requirement for different ancestors

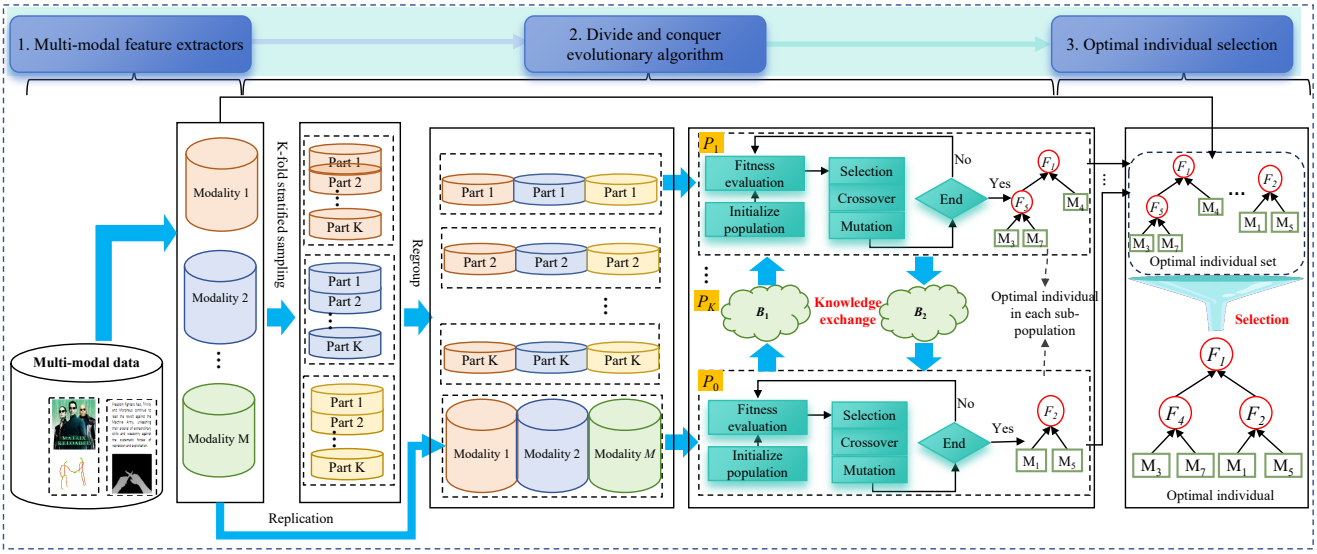

Figure 1: The whole framework of DC-NAS.

in each unit, leading to limited fusion options and results in a decrease in performance. Moreover, BM-NAS (Yin et al. 2022) adoption of tensor representation for modal feature fusion introduces additional computational overhead and information redundancy. MMnas (Yu et al. 2020) allows the search of attention operations, but the network's topology remains fixed during architecture search.

Diverging from the aforementioned approaches, the proposed DC-NAS is an evolution-based NAS approach with a wider and more flexible search space than gradient-based methods, achieving superior architectural performance. Moreover, DC-NAS utilizes feature vectors as the foundational fusion units, exclusively uses five fundamental fusion operators for feature fusion and uses the divide-and-conquer algorithm for population evolution, significantly reducing computational load and addressing the low efficiency issue of the existing population-based NAS methods.

The Proposed DC-NAS

For avoiding confusion, we appoint some terms here. The population consists of individuals, each individual corresponds to a multi-modal classification model that is encoded as a tree. The knowledge extracted from the sub-populations typically refers to the special individuals or parts of them. The all representations extracted from modalities are uniformly called features.

In this paper, we propose an efficient evolutionary-based NAS method for multi-model classification called divide-and-conquer neural architecture search (DC-NAS). DC-NAS efficiently searches for optimal DNN architectures that fuse multi-modal features. It employs an evolutionary algorithm to iteratively improve and optimize the fusion architectures in the population. Throughout the population iterations, we adopt the divide-and-conquer strategy to partition the training dataset into multiple disjoint subsets and allocate each subset to a separate sub-population. Additionally, a special sub-population is evolved on the entire dataset for obtaining more accuracy individuals. Knowledge exchange between each sub-population and one of two knowledge bases that is achieved via the crossover operator enables effective knowledge transfer among sub-populations, ensuring learning performance. Figure 1 illustrates the entire DC-NAS framework.

Unimodal Feature Extraction

In this study, we follow previous works on multi-modal fusion such as MFAS(Perez Rua et al. 2019), MMTM(Vaezi Joze et al. 2020), and BM-NAS(Yin et al. 2022) to adopt the pre-trained unimodal neural network models as feature extractors. We extract raw features from the intermediate layers of these models, as neural network architectures typically have a layered or block-like structure, which naturally lends itself to this extraction approach. Since the features extracted from different modalities have varying dimensions (e.g., one-dimensional for text and two-dimensional for images), we employ global average pooling to transform them into feature vectors, achieving feature alignment and facilitating subsequent feature fusion while also reducing computational complexity. By extracting intermediate layers from a multi-modal single neural network, we can obtain a collection of n modal feature datasets represented by the dataset X, where $X = \{(X_1(s_i), X_2(s_i), X_v(s_i), y_i)\}_{i=1}^{n}$. Here, $X_j(s_i)$ represents the j-th feature of the multi-modal data s_i, and X_j represents the j-th feature representation extracted from the multi-modal dataset.

Multi-Modal Classification Model Encoding and Encoding

The each individual p in population is encoded as a binary tree, where the leaf nodes consist of features and the branch nodes consist of fusion operators. In this paper, the fusion

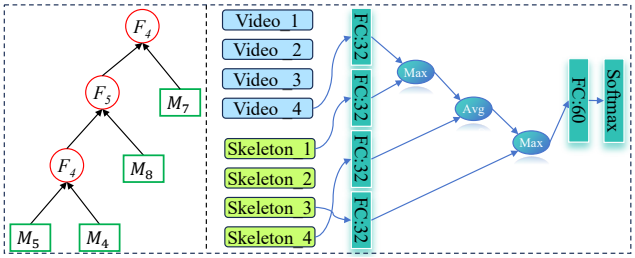

Figure 2: Left: the binary tree encoding of the optimal individual for NTU RGB-D dataset. Right: its corresponding multi-modal classification model.

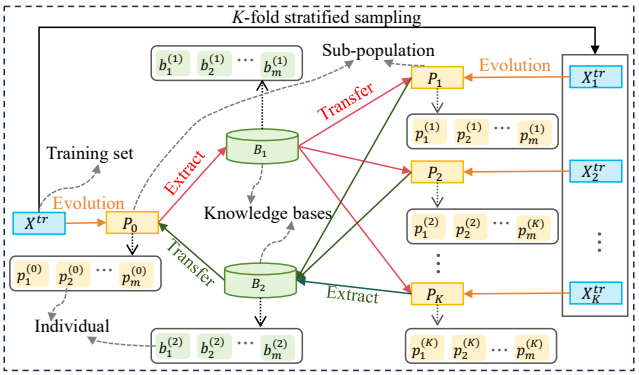

Figure 3: The relationship among different variables.

operators consists of concatenation (Wang et al. 2017), addition (Wu et al. 2014), multiplication (Gao et al. 2018), Maximize(Duong, Lebret, and Aberer 2017) and average. Their definitions can be found in appendix. For each individual, if the binary tree contains k features, then it must contain $k-1$ fusion operators. Each individual corresponds to multi-modal classification model. The left in Figure 2 shows the optimal individual on NTU RGB-D dataset (Shahroudy et al. 2016). The binary tree can be decoded a multi-modal classification model shown in the right in Figure 2 as following steps: 1) Pass the modality features represented by the leaf nodes of the individual encoding tree into fully connected layers (FC) for feature alignment to facilitate feature fusion; 2) Perform feature fusion based on the fusion operators represented by the branch nodes; 3) Pass the fused features into a FC and a Softmax layer for the final prediction output.

Algorithm Framework

The core idea of DC-NAS is that the individuals in population are trained using partial training set, instead of entire one. As shown in Figure 3, the detailed process of our divide-and-conquer approach and the relationships between various variables are presented. Specifically, the K-fold stratified sampling is first conducted on the entire training dataset X^{tr}, obtaining K non-overlapping subsets with the same class proportion, denoted as $X_1^{tr}, X_2^{tr}, ..., X_K^{tr}$; the entire population P with M individuals is evenly divided into $K+1$ sub-populations, namely $P_0, P_1, ..., P_K$. Each sub-population contains m individuals, where $m = \lfloor M/(K+1) \rfloor$. For instance, when we perform 3-fold stratified sampling on the training set, the population with 28 individuals will be divided into four sub-populations, each has seven individuals. The sub-population P_0 will evolves on the entire dataset X^{tr}, while P_i will evolves on the subset X_i^{tr}, where $i = 1, 2, \cdots, K$.

Each sub-population P_i in DC-NAS can learn different knowledge from distinct training data. To enable fast knowledge transfer among the $K+1$ sub-populations at each generation, we propose a simple yet effective knowledge transfer method. Specifically, given two knowledge bases B_1 and B_2 that store the m best optimal individuals of P_0 and $\{P_1, ..., P_K\}$ of all the past generations, respectively. The storage of each knowledge base is limited to m to conserve memory. For example, suppose that the current population generation is t and P_i^t denotes the offspring of the sub-population P_i in t-th generation. Then, B_1 stores the top-m best optimal individuals extracted from $\{P_0^1, P_0^2, \cdots, P_0^t\}$; while B_2 stores the top-m best ones that are obtained by extracting m/K from each $\{P_i^1, P_i^2, \cdots, P_i^t\}$, $i = 1, 2, \cdots, K$, respectively. After fitness evaluation in each generation, both knowledge bases B_1 and B_2 will be updated. Knowledge exchange is performed via crossover, the details will be described in Crossover part.

We employ a standard evolutionary algorithm to search for the optimal solution through population-based evolutionary NAS. The main steps of the DC-NAS framework include population initialization, fitness evaluation, offspring generation, and selection.

Population Initialization: A population P with M individuals is randomly generated, and then divide it into $K+1$ sub-populations.

Fitness Evaluation: Each individual is first decoded into a multi-modal classification model, and then the model is trained using the corresponding to sub-dataset. Its classification accuracy or weighted F1 on test set is used as the fitness value for the decoded individual.

Crossover, Mutation, and Selection: As shown in Figure 4, there exists two cases for crossover. The first case involves the crossover between individuals from P_0 and B_2, while the second case involves the crossover between individuals from $P_1, P_2, ..., P_K$ and B_1. Let's take the first case as an example: we first select an individual from P_0 and another individual from B_2 for crossover. This process is repeated $m/2$ times. After the crossover, the population undergoes mutation processing, followed by binary tournament selection (Miller and Goldberg 1996) involving both offspring and parents. The pseudo code of DC-NAS is shown in Algorithm 1.

Experiments

In this study, we evaluated DC-NAS on three popular multi-modal tasks: (1) multi-label movie genre classification task on the MM-IMDB dataset (Arevalo et al. 2017), (2) multi-modal action recognition task on the NTU RGB-D dataset (Shahroudy et al. 2016), and (3) multi-modal gesture recognition task on the EgoGesture dataset (Zhang et al. 2018).

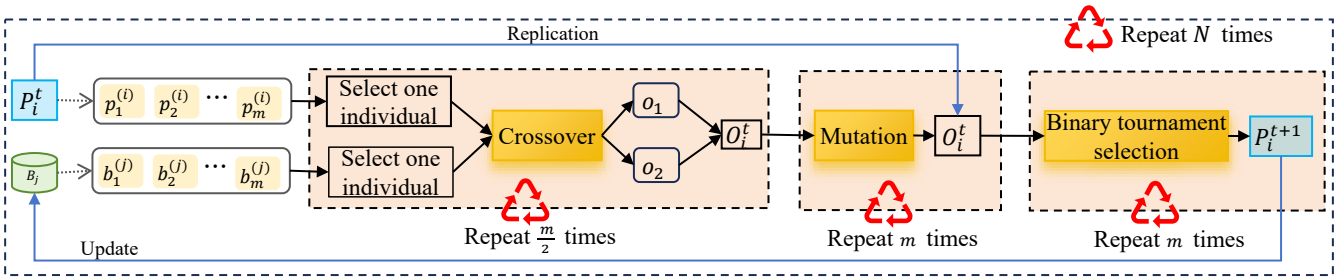

Figure 4: Evolution process of each sub-population P_i^t and each knowledge base B_j where $i \in \{0, 1, \cdots, K\}$, $j \in \{1, 2\}$, and $t \in \{1, 2, \cdots, N\}$. When i takes 0, j is equal to 2; When i takes any values from $\{1, 2, \cdots, K\}$, j is equal to 1.

For each task, we conducted a brief analysis and provided the experimental parameter settings.

MM-IMDB Dataset

The MM-IMDB dataset (Arevalo et al. 2017) is a comprehensive multi-modal collection sourced from the Internet Movie Database, encompassing 25,959 movies along with their associated posters, plots, genres, and other metadata. In our experiment, we focus on the task of multi-label genre classification, utilizing both posters (RGB images) and plots (text) as input modalities. The dataset includes a total of 27 non-mutually exclusive genres, such as Drama, Comedy, Romance, and more. However, due to severe class imbalance, we opt to employ only 23 genres for the classification task, omitting News, Adult, Talk-Show, and Reality-TV genres as they account for just 0.10% of the dataset. The dataset is originally split into three subsets: 15,552 movies for training, 2,608 for validation, and 7,799 for testing purposes. The processing approach is also designed to facilitate a comparison with previous methods.

To ensure a fair comparison with other explicit multi-modal fusion methods, the same neural network backbone models as BM-NAS (Yin et al. 2022) are adopted in our experiments. Specifically, Maxout MLP (Goodfellow et al. 2013) is chosen as the backbone model of the text modality, and VGG Transfer (Simonyan and Zisserman 2015), a deep neural network model based on VGG19 (Simonyan and Zisserman 2015), is chosen as the backbone model of the RGB image modality. The evaluation metric is the weighted F1 score, which is a reliable measure of multi-label classification performance due to the high degree of imbalance in the dataset, rather than other types of F1 scores. The use of weighted F1 scores is also consistent with previous approaches to facilitate comparison. For the parameters of our architecture, we set the population size N = 20, the number of population iterations T = 10, the dimension of the fusion vector FD = 256 and the modal features are repeatable.

Due to the highly imbalanced class distribution in the MM-IMDB dataset (Arevalo et al. 2017), we adopt a more reasonable metric, the weighted F1 score, to measure the performance of multi-label classification. As shown in Table 1, compared to existing multi-modal classification methods, DC-NAS achieves the best weighted F1 score, outperforming the latest BM-NAS (Yin et al. 2022) method by 0.78%.

Method	Modality	F1-W(%)
Unimodal methods		
Maxout MLP (ICML13)	Text	57.54
VGG Transfer (ICLR15)	Image	49.21
Multi-modal methods		
Two-stream (NIPS14)	Image + Text	60.81
GMU (ICLR17)	Image + Text	61.70
CentralNet (ECCV18)	Image + Text	62.23
MFAS (CVPR19)	Image + Text	62.50
BM-NAS (AAAI22)	Image + Text	62.92±0.03
DC-NAS (ours)	Image + Text	**63.70**±0.11

Table 1: Multi-label genre classification results on MM-IMDB dataset. Weighted F1 (F1-W) is reported.

Method	Modality	Acc (%)
Unimodal methods		
Inflated ResNet-50 (CVPR18)	Video	83.91
Co-occurence (IJCAI18)	Pose	85.24
Multi-modal methods		
Two-stream (NIPS14)	Video + Pose	88.60
GMU (ICLR17)	Video + Pose	85.80
MMTM (CVPR20)	Video + Pose	88.92
CentralNet (ECCV18)	Video + Pose	89.36
MFAS (CVPR19)	Video + Pose	89.50±0.60
BM-NAS (AAAI22)	Video + Pose	90.48±0.24
DC-NAS (ours)	Video + Pose	**90.85±0.05**

Table 2: Action recognition results on NTU RGB-D dataset.

NTU RGB-D Dataset

A large-scale multi-modal action recognition dataset from NTU RGB-D (Shahroudy et al. 2016) consists of 56,880 samples representing 40 subjects, 80 viewpoints, and 60 daily activities. For our fusion experiments, we utilize skeleton and RGB video modalities. The performance evaluation employs the Cross-Subject (CS) accuracy metric. As a result of maintaining consistency, we adopt the dataset split introduced in BM-NAS (Yin et al. 2022), in which subjects 1, 4, 8, 13, 15, 17, and 19 are used for training, subjects 2, 5, 9, and 14 are used for validation, and the remaining subjects

Algorithm 1: Divide-and-conquer neural architecture search (DC-NAS)

Input: Training data X^{tr}, test dataset X^{te}, number of sub-populations $K+1$.
Parameter: Population size M, maximum number of generations N.
Output: Optimal fusion individual.

Initialize $t \leftarrow 1$;
Initialize two empty knowledge bases $B_1 \leftarrow []$ and $B_2 \leftarrow []$;
Initialize a population P with M individuals;
Divide the population P into $K+1$ sub-populations: $P_0, P_1, ..., P_K$;
Conduct K-fold stratified sampling on X^{tr}, obtain $X_1^{tr}, X_2^{tr}, ..., X_K^{tr}$;
P_0 gets X^{tr}, while P_i get X_i^{tr}, $i = 1, 2, \cdots, K$;
Train and evaluate each individual in each sub-population;
Select the best fusion individual from each sub-population and add it to the corresponding knowledge base;
while $t <= N$ **do**
    Generate offspring $Q_0^t, Q_1^t, ..., Q_K^t$ using the crossover operator, where Q_0^t is produced by crossing P_0^t and B_2^t, $Q_1^t, Q_2^t, ..., Q_K^t$ are produced by crossing $P_1^t, P_2^t, ..., P_K^t$ with B_1^t;
    Conduct mutation on each individual in Q_j^t, where $j = 0, 1, 2, ..., K$;
    Train and evaluate each individual in each sub-population;
    Select next generation population P_i^{t+1} from $Q_i^t \cup P_i^t$, where $i = 0, 1, 2, ..., K$;
    Select the best fusion individual from each sub-population and add it to the corresponding knowledge base;
    Update the knowledge bases: Knowledge Bank1 B_1^t, Knowledge Bank2 B_2^t;
    $t = t + 1$;
end while

are used for testing. The training, validation, and testing sets comprise 23,760, 2,519, and 16,558 samples, respectively.

For a fair comparison, two convolutional neural network models are used as the modal feature extractors, following the same approach as the BM-NAS (Yin et al. 2022) method. Specifically, Inflated ResNet-50 (Baradel et al. 2018) is employed for the video modality and Co-occurrence (Li et al. 2018) for the skeleton modality. This design ensures that all methods in the experiment share the same backbone network. Additionally, we follow the data preprocessing pipeline of BM-NAS (Yin et al. 2022), MFAS(Perez Rua et al. 2019) and MMTM (Vaezi Joze et al. 2020) to ensure the fairness of the experimental results. We use a population size of 28, conduct 15 iterations, do not reuse modalities, and set the fusion modality dimension to be 64.

In Table 2, our method achieve a cross-subject accuracy of 90.85%, demonstrating superior results compared to the latest approaches on NTU RGB-D (Shahroudy et al. 2016) using video and pose modalities. When compared to the recent BM-NAS (Yin et al. 2022) framework, our DC-NAS has several advantages. BM-NAS (Yin et al. 2022) employs tensor-based fusion, while DC-NAS adopts a vector-based approach. Although tensor fusion provides more information, it also introduces redundancy and increases computational overhead, leading to performance degradation. Additionally, it may disrupt the structural information of feature vectors generated by the penultimate fully connected layer. Furthermore, BM-NAS (Yin et al. 2022) relies on gradient-based NAS, whereas our DC-NAS utilizes evolution-based NAS. Evolution-based methods typically explore a larger space of modality fusion and fusion strategies, making it more likely to discover better solutions.

Method	Modality	Acc (%)
Unimodal methods		
VGG-16 + LSTM (NIPS14)	RGB	74.70
C3D + LSTM + RSTTM	RGB	89.30
I3D (CVPR17)	RGB	90.33
ResNext-101 (FG19)	RGB	93.75
VGG-16 + LSTM (CVPR14)	Depth	77.70
C3D + LSTM + RSTTM	Depth	90.60
I3D (CVPR17)	Depth	89.47
ResNeXt-101 (FG19)	Depth	94.03
Multi-modal methods		
VGG-16 + LSTM (CVPR17)	RGB + Depth	81.40
C3D + LSTM + RSTTM	RGB + Depth	92.20
I3D (CVPR17)	RGB + Depth	92.78
MMTM (CVPR20)	RGB + Depth	93.51
MTUT (3DV19)	RGB + Depth	93.87
3D-CDC-NAS2 (TIP21)	RGB + Depth	94.38
BM-NAS (AAAI22)	RGB + Depth	94.96±0.07
DC-NAS (ours)	RGB + Depth	**95.22±0.05**

Table 3: Gesture recognition results on EgoGesture dataset.

Method	Dataset	Parameters	Time	CP (%)
MMTM	NTU	8.61M	-	88.92
MFAS	NTU	2.16M	603.64	89.50
BM-NAS	NTU	0.98M	53.68	90.48
DC-NAS(ours)	NTU	**0.26M**	**13.63**	**90.85**
BM-NAS	Ego	0.61M	20.67	94.96
DC-NAS(ours)	Ego	**0.19M**	**4.57**	**95.22**
BM-NAS	MM-IMDB	0.65M	1.24	62.94
DC-NAS(ours)	MM-IMDB	**0.42M**	**1.19**	**63.70**

Table 4: Comparison of model size, search cost (GPU hours), and classification performance (CP) of generalized multi-modal NAS methods.

EgoGesture Dataset

The EgoGesture dataset (Zhang et al. 2018) comprises a large-scale multi-modal gesture recognition dataset, with 24,161 gesture samples gathered from 50 diverse subjects and 6 distinct scenes, encompassing a total of 83 unique gesture categories. To maintain experimental fairness, we

Version	DCE	KT	Time	ACC (%)
DC-NAS$_1$	False	False	20.67	90.86±0.03[8.0e-01]
DC-NAS$_2$	True	False	11.10	90.52±0.06[9.5e-06]
DC-NAS	True	True	13.63	90.85±0.05

Table 5: Ablation study on the two core component of DC-NAS on NTU RGB-D dataset. $[\cdot]$ shows the p values of the paired t-test between DC-NAS and other two versions.

adhered to the dataset's original division, wherein samples were grouped based on subjects. Specifically, our training set consisted of 14,416 samples, the validation set had 4,768 samples, and the testing set comprised 4,977 samples.

For a fair comparison, we follow the setup of the BM-NAS (Yin et al. 2022) method and utilize ResNeXt-101 (Kpkl et al. 2019) as the backbone for RGB and depth video modalities. We compare our DC-NAS with various single-modal and multi-modal methods. The experimental settings for DC-NAS involves a population size of 28, 15 iterations, and non-reuse of modalities with a fusion dimension of 32. Table 3 presents the experimental results on the EgoGesture dataset. Compared to other methods, DC-NAS achieves the state-of-the-art classification performance.

Comparison of Searching Time and Model Size

This section aims to show the advantage of DC-NAS by comparing it with three strong MMC benchmark methods including MFAS (Perez Rua et al. 2019), BM-NAS (Yin et al. 2022) and MMTM (Vaezi Joze et al. 2020) in terms of the searching time, model size and classification performance. MMTM is a manual MMC methods and the rest ones are NAS methods. The results are reported in Table 4. From the Table 4, we observe that our DC-NAS costs the least searching time, but finds the optimal model with less parameters and better classification performance. For example, the model size of DC-NAS is reduced by at least three times on the NTU RGB-D and EgoGesture; The time consumption for searching the optimal fusion model on the NTU RGB-D and EgoGesture datasets has been nearly reduced by four times than the state-of-the-art BM-NAS. These results furthermore demonstrate the effectiveness of DC-NAS.

Ablation Study

The section aims to verify the effectiveness of each component of DC-NAS, the unimodal feature selection strategy and the multi-modal fusion strategy by three ablation studies.

Impact Analysis of Each Component of DC-NAS. DC-NAS includes two core component: divide-and-conquer evolution (DCE) and knowledge transfer (KT). This section aims to evaluate the importance of each component by comparing DC-NAS with its two modified versions DC-NAS$_1$ and DC-NAS$_2$ in Table 5. Specifically, DC-NAS$_1$ denotes all individuals are trained using the entire training dataset; DC-NAS$_2$ is the version of DC-NAS without KT module. From the Table 5, there is no significant difference between DC-NAS and DC-NAS$_1$ according to the paired t-test with a 95% confidence level. However, DC-NAS is significantly better than DC-NAS$_2$ according to the paired t-test with a 95% confidence level. These results indicate that jointly using them ensures that training individuals using partial data can achieve almost the same classification performance.

Feature selection strategies	ACC (%)
Random	88.81±0.11
Late fusion	89.47±0.07
Searched (MFAS)	89.50±0.60
Searched (BM-NAS)	90.48±0.24
Searched (DC-NAS)	**90.85±0.05**

Table 6: Ablation study for feature selection strategies on NTU RGB-D Dataset.

Add	Mul	Cat	Max	Avg	DC-NAS
89.54	88.71	89.20	88.84	88.07	**90.85**

Table 7: Ablative study for fusion operators on NTU RGB-D dataset.

Impact Analysis on Feature Selection Strategies. Table 6 compares different unimodal feature selection strategies on NTU RGB-D dataset. For the random strategy, we evaluate the average results of the first 10 individuals initialized randomly in the population. As for the late fusion strategy, we concatenate the last two layers of features extracted from two modalities, train and evaluate the model five times, and then take the average value. We then compare these strategies with the MFAS (Perez Rua et al. 2019) and BM-NAS (Yin et al. 2022) methods for selecting modal fusion strategies. As shown in Table 6, the searched feature selection strategy is better than all baselines, demonstrating that feature selection strategy plays an important role in multi-modal classification task.

Impact Analysis on Fusion Strategy. Table 7 evaluates different multimodal fusion strategies on NTU RGB-D dataset. We compare the optimal feature fusion individuals found by DC-NAS with the direct use of all modality features using the five basic fusion operators. As shown in Table 7, the searched fusion strategy is better than all baselines, demonstrating that the optimal fusion strategy of different features may be different.

Conclusion

In this paper, we have proposed an efficient NAS-MMC method DC-NAS. DC-NAS reduces the searching time by training different sub-populations using small-scale data, while achieves the comparable or even better classification performance by exchanging knowledge between sub-populations. Extensive experiments have been conducted for verifying its these advantages. DC-NAS may make it possible that NAS-MMC technique applies to the large-scale multi-modal data. In the future, some important issues related to DC-NAS need to be studied more deeply, such as more effectively strategies of the knowledge exchange between different sub-populations and data split.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2021ZD0112400), National Natural Science Foundation of China (Nos. 62136005, 62306171, 61976129, 62106132, 62306170, 61906115), the Science and Technology Major Project of Shanxi (No. 202201020101006), Young Scientists Fund of the Natural Science Foundation of Shanxi (Nos. 202203021222183, 20210302124549), Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province (Nos. CICIP2023005, CICIP202205), Science and Technology Innovation Plan for Colleges and Universities of Shanxi Province (2022L296), and Taiyuan University of Science and Technology Doctoral Research Start-up Fund Project(20222106).

References

Arevalo, J.; Solorio, T.; Montes-y Gmez, M.; and Gonzlez, F. 2017. Gated Multimodal Units for Information Fusion. *Cornell University - arXiv*.

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 469–478.

Bi, Y.; Xue, B.; and Zhang, M. 2021. A Divide-and-Conquer Genetic Programming Algorithm With Ensembles for Image Classification. *IEEE Transactions on Evolutionary Computation*, 25(6): 1148–1162.

Dong, X.; Kedziora, D. J.; Musial, K.; and Gabrys, B. 2021. Automated Deep Learning: Neural Architecture Search Is Not the End. *ArXiv*, abs/2112.09245.

Duong, C. T.; Lebret, R.; and Aberer, K. 2017. Multimodal Classification for Analysing Social Media. *ArXiv*, abs/1708.02099.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.; Wang, X.; and Li, H. 2018. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6632–6641.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; and Bengio, Y. 2013. Maxout Networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28 of *ICML'13*, 13191327. JMLR.org.

Guo, Q.; Qian, Y.; and Liang, X. 2022. GLRM: Logical Pattern Mining in the Case of Inconsistent Data Distribution based on Multigranulation Strategy. *International Journal of Approximate Reasoning*, 143: 78–101.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2023. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2551–2566.

Hou, M.; Tang, J.; Zhang, J.; Kong, W.; and Zhao, Q. 2019. Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. In *Proceedings of the 32th Advances in Neural Information Processing Systems*, 12136–12145.

Jiang, B.; Zhang, C.; Zhong, Y.; Liu, Y.; Zhang, Y.; Wu, X.; and Sheng, W. 2023. Adaptive Collaborative Fusion for Multi-View Semi-Supervised Classification. *Information Fusion*, 96: 37–50.

Kpkl, O.; Gunduz, A.; Kose, N.; and Rigoll, G. 2019. Real-Time Hand Gesture Detection and Classification Using Convolutional Neural Networks. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition*, 1–8.

Lardeux, F.; and Goffon, A. 2010. *A Aynamic Island-based Genetic Algorithms Framework*, 156165.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018. Co-Occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 786792. AAAI Press.

Liang, X.; Guo, Q.; Qian, Y.; Ding, W.; and Zhang, Q. 2021. Evolutionary Deep Fusion Method and Its Application in Chemical Structure Recognition. *IEEE Transactions on Evolutionary Computation*, 25(5): 883–893.

Liang, X.; Qian, Y.; Guo, Q.; Cheng, H.; and Liang, J. 2022. AF: An Association-Based Fusion Method for Multi-Modal Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9236–9254.

Liang, X.; Qian, Y.; Guo, Q.; and Zheng, K. 2023. A Data Representation Method Using Distance Correlation. *Frontiers of Computer Science*, 1–16.

Liu, H.; Simonyan, K.; and Yang, Y. 2019. DARTS: Differentiable Architecture Search. In *Proceedings of the International Conference on Learning Representations*, 1–11.

Liu, W.; Chen, Y.; Yue, X.; Zhang, C.; and Xie, S. 2023. Safe Multi-View Deep Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8870–8878.

Miller, B. L.; and Goldberg, D. E. 1996. Genetic Algorithms, Selection Schemes, and the Varying Effects of Noise. *Evolutionary Computation*, 4(2): 113–131.

Peng, Y.; Bi, L.; Fulham, M.; Feng, D.; and Kim, J. 2020. Multi-Modality Information Fusion for Radiomics-based Neural Architecture Search. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, 763–771. Cham: Springer International Publishing. ISBN 978-3-030-59728-3.

Perez Rua, J.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. MFAS: Multimodal Fusion Architecture Search. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6959–6968.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Vaezi Joze, H. R.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: Multimodal Transfer Module for CNN Fusion. In *Proceedings of the 2020 IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition, 13286–13296.

Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2019. CentralNet: A Multilayer Approach for Multimodal Fusion. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision – ECCV 2018 Workshops*, 575–589. Springer International Publishing. ISBN 978-3-030-11024-6.

Wang, L.; Li, W.; Li, W.; and Gool, L. V. 2017. Appearance-and-Relation Networks for Video Classification. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1430–1439.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep Double Incomplete Multi-View Multi-Label Learning With Incomplete Labels and Missing Views. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Wu, Z.; Jiang, Y.-G.; Wang, J.; Pu, J.; and Xue, X. 2014. Exploring Inter-Feature and Inter-Class Relationships with Deep Neural Networks for Video Classification. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, 167176.

Xu, C.; Zhao, J.; Guan, Z.; Yang, Y.; Chen, L.; and Song, X. 2023a. Progressive Deep Multi-View Comprehensive Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10557–10565.

Xu, J.; Li, C.; Peng, L.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023b. Adaptive Feature Projection With Distribution Alignment for Deep Incomplete Multi-View Clustering. *IEEE Transactions on Image Processing*, 32: 1354–1366.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Yu, P. S.; and He, L. 2023c. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7470–7482.

Yin, Y.; Huang, S.; Zhang, X.; and Dou, D. 2022. BM-NAS: Bilevel Multimodal Neural Architecture Search. In *Association for the Advancement of Artificial Intelligence*, 8901–8909.

Yu, Z.; Cui, Y.; Yu, J.; Wang, M.; Tao, D.; and Tian, Q. 2020. Deep Multimodal Neural Architecture Search. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 37433752. New York, NY, USA: Association for Computing Machinery.

Yu, Z.; Zhou, B.; Wan, J.; Wang, P.; Chen, H.; Liu, X.; Li, S. Z.; and Zhao, G. 2021. Searching Multi-Rate and Multi-Modal Temporal Enhanced Networks for Gesture Recognition. *IEEE Transactions on Image Processing*, 30: 5626–5640.

Yuan, G.; Wang, B.; Xue, B.; and Zhang, M. 2023. Particle Swarm Optimization for Efficiently Evolving Deep Convolutional Neural Networks Using an Autoencoder-based Encoding Strategy. *IEEE Transactions on Evolutionary Computation*.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2022. Deep Partial Multi-View Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2402–2415.

Zhang, Y.; Cao, C.; Cheng, J.; and Lu, H. 2018. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Transactions on Multimedia*, 10381050.

Zhou, D.; Zhou, D.; Hu, D.; Zhou, H.; Bai, L.; Liu, Z.; and Ouyang, W. 2022. SepFusion: Finding Optimal Fusion Structures for Visual Sound Separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3544–3552.

Zhuge, W.; Tao, H.; Luo, T.; Zeng, L.-L.; Hou, C.; and Yi, D. 2022. Joint Representation Learning and Clustering: A Framework for Grouping Partial Multiview Data. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3826–3840.

Zoph, B.; and Le, Q. V. 2017. Neural Architecture Search with Reinforcement Learning. In *5th International Conference on Learning Representations*, 1–14.