The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

BM-NAS: Bilevel Multimodal Neural Architecture Search

Yihang Yin[1], **Siyu Huang**[2], **Xiang Zhang**[3]

[1]Nanyang Technological University
[2]Harvard University
[3]The Pennsylvania State University
yyin009@e.ntu.edu.sg, huang@seas.harvard.edu, xzz89@psu.edu

Abstract

Deep neural networks (DNNs) have shown superior performances on various multimodal learning problems. However, it often requires huge efforts to adapt DNNs to individual multimodal tasks by manually engineering unimodal features and designing multimodal feature fusion strategies. This paper proposes Bilevel Multimodal Neural Architecture Search (BM-NAS) framework, which makes the architecture of multimodal fusion models fully searchable via a bilevel searching scheme. At the upper level, BM-NAS selects the inter/intra-modal feature pairs from the pretrained unimodal backbones. At the lower level, BM-NAS learns the fusion strategy for each feature pair, which is a combination of predefined primitive operations. The primitive operations are elaborately designed and they can be flexibly combined to accommodate various effective feature fusion modules such as multi-head attention (Transformer) and Attention on Attention (AoA). Experimental results on three multimodal tasks demonstrate the effectiveness and efficiency of the proposed BM-NAS framework. BM-NAS achieves competitive performances with much less search time and fewer model parameters in comparison with the existing generalized multimodal NAS methods. Our code is available at https://github.com/Somedaywilldo/BM-NAS.

Introduction

Deep neural networks (DNNs) have achieved a great success on various unimodal tasks (*e.g.*, image categorization (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016), language modeling (Vaswani et al. 2017; Devlin et al. 2018), and speech recognition (Amodei et al. 2016)) as well as the multimodal tasks (*e.g.*, action recognition (Simonyan and Zisserman 2014; Vielzeuf et al. 2018), image/video captioning (You et al. 2016; Jin et al. 2019, 2020; Yan et al. 2021), visual question answering (Lu et al. 2016; Anderson et al. 2018), and cross-modal generation (Reed et al. 2016; Zhou et al. 2019)). Despite the superior performances achieved by DNNs on these tasks, it usually requires huge efforts to adapt DNNs to the specific tasks. Especially with the increase of modalities, it is exhausting to manually design the backbone architectures and the feature fusion strategies. It raises urgent concerns about the automatic design of multimodal DNNs with minimal human interventions.

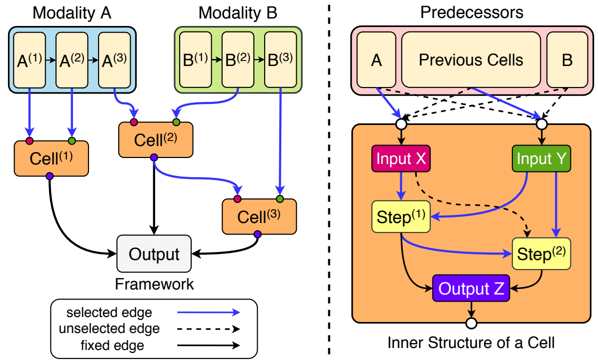

Figure 1: An overview of our BM-NAS framework for multimodal learning. a *Cell* is a searched feature fusion unit that accepts two inputs from modality features or other Cells. In a bilevel fashion, we search the connections between Cells and the inner structures of Cells, simultaneously.

Neural architecture search (NAS) (Zoph and Le 2017; Liu et al. 2018a) is a promising data-driven solution to this concern by searching for the optimal neural network architecture from a predefined space. By applying NAS to multimodal learning, MMnas (Yu et al. 2020) searches the architecture of Transformer model for visual-text alignment and MMIF (Peng et al. 2020) searches the optimal CNNs structure to extract multi-modality image features for tomography. These methods lack generalization ability since they are designed for models on specific modalities. MFAS (Pérez-Rúa et al. 2019) is a more generalized framework which searches the feature fusion strategy based on the unimodal features. However, MFAS only allows fusion of inter-modal features, and the fusion operations are not searchable. It results in a limited space of feature fusion strategies when dealing with various modalities in different multimodal tasks.

In this paper, we propose a generalized framework, named Bilevel Multimodal Neural Architecture Search (BM-NAS), to adaptively learn the architectures of DNNs for a variety of multimodal tasks. BM-NAS adopts a *bilevel* searching scheme that it learns the unimodal feature selection strategy at the upper level and the multimodal feature fusion strategy at the lower level, respectively. As shown in the

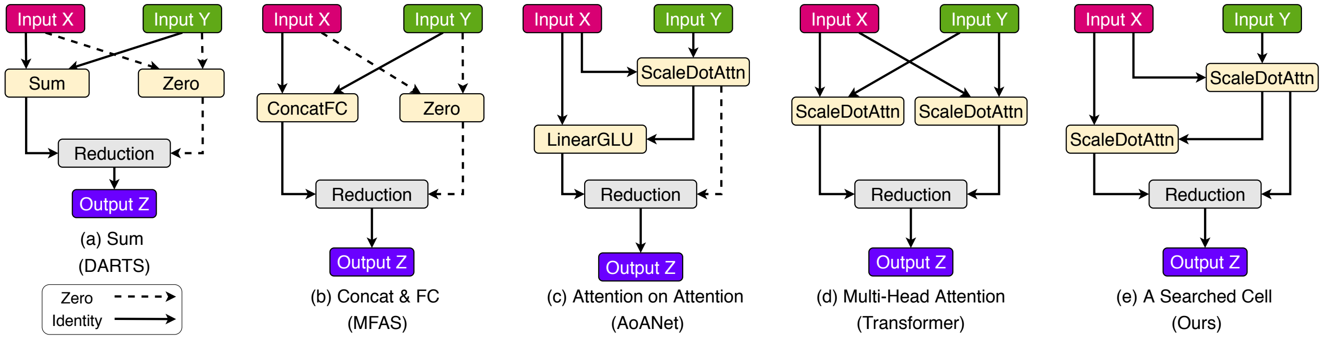

Figure 2: The search space of a Cell in BM-NAS accommodates many existing multimodal fusion strategies. (d) is a two-head version of multi-head attention (Vaswani et al. 2017), and more heads can be flexibly added by changing the number of inner steps. (e) is the Cell founded by BM-NAS on NTU RGB-D dataset (Shahroudy et al. 2016), and it outperforms these existing fusion strategies (see Table 7).

left part of Fig. 1, the upper level of BM-NAS consists of a series of feature fusion units, *i.e.*, Cells. The Cells are organized to combine and transform the unimodal features to the task output through a searchable directed acyclic graph (DAG). The right part of Fig. 1 illustrates the lower level of BM-NAS which learns the inner structures of Cells. A Cell is comprised of several predefined primitive operations. We carefully select the primitive operations such that different combinations of them can form a large variety of fusion modules, as shown in Fig. 2, our search space incorporates benchmark attention mechanisms like multi-head attention (Transformer) (Vaswani et al. 2017) and Attention on Attention (AoA) (Huang et al. 2019). The bilevel scheme of BM-NAS is end-to-end learned using the differentiable NAS framework (Liu, Simonyan, and Yang 2019). We conduct extensive experiments on three multimodal tasks to evaluate the proposed BM-NAS framework. BM-NAS shows superior performances in comparison with the state-of-the-art multimodal methods. Compared with the existing generalized multimodal NAS frameworks, BM-NAS achieves competitive performances with much less search time and fewer model parameters. To the best of our knowledge, BM-NAS is the first multimodal NAS framework that supports the search of both the unimodal feature selection strategies and the multimodal fusion strategies.

The main contributions of this paper are three-fold.

1. Towards a more generalized and flexible design of DNNs for multimodal learning, we propose a new paradigm that employs NAS to search both the unimodal feature selection strategy and the multimodal fusion strategy.
2. We present a novel BM-NAS framework to address the proposed paradigm. BM-NAS makes the architecture of multimodal fusion models fully searchable via a bilevel searching scheme.
3. We conduct extensive experiments on three multimodal learning tasks to evaluate the proposed BM-NAS framework. Empirical evidences indicate that both the unimodal feature selection strategy and the multimodal fusion method are significant to the performance of multimodal DNNs.

Related Work

Neural Architecture Search

Neural architecture search (NAS) aims at automatically finding the optimal neural network architectures for specific learning tasks. NAS can be viewed as a bilevel optimization problem by optimizing the weights and the architecture of DNNs at the same time. Since the network architecture is discrete, traditional NAS methods usually rely on the black-box optimization algorithms, resulting in a extremely large computing cost. For example, searching architectures using reinforcement learning (Zoph and Le 2016) or evolution (Real et al. 2019) would require thousands of GPU-days to find a state-of-the-art architecture on ImageNet dataset (Deng et al. 2009) due to low sampling efficiency.

As a result, many methods were proposed for speeding up NAS. From the perspective of engineering, ENAS (Pham et al. 2018) improve the sampling efficiency by weight-sharing. From the perspective of optimization algorithm, PNAS (Liu et al. 2018b) employs sequential model-based optimization (SMBO) (Hutter, Hoos, and Leyton-Brown 2011), using surrogate model to predict the performance of an architecture. Monte Carlo tree search (MTCS) (Negrinho and Gordon 2017) and Bayesian optimization (BO) (Kandasamy et al. 2018) are also explored to enhance the sampling efficiency.

Recently, a remarkable efficiency improvement of NAS is achieved by differentiable architecture search (DARTS) (Liu, Simonyan, and Yang 2019). DARTS introduces a continuous relaxation of the network architecture, making it possible to search an architecture via gradient-based optimization. However, DARTS only supports the search of unary operations. For specific multimodal tasks, we expect the NAS framework to support the search of multi-input operations, in order to obtain the optimal fusion strategy. In this work, we devise a novel NAS framework named BM-NAS for multimodal learning. BM-NAS follows the optimization scheme of DARTS, however, it novelly introduces a bilevel searching scheme to search the unimodal feature selection strategy and the multimodal fusion strategy simultaneously, enabling an effective search scheme for multimodal fusion.

Multimodal Fusion

The multimodal fusion techniques for DNNs can be generally classified into two categories: early fusion and late fusion. Early fusion combines low-level features, while late fusion combines prediction-level features. To combine these features, a series of reduction operations such as weighted average (Natarajan et al. 2012) and bilinear product (Teney et al. 2018) are proposed in previous works. As each unimodal DNNs backbone could have tens of layers or maybe more, manually sorting out the best intermediate features for multimodal fusion could be exhausting. Therefore, some works propose to enable fusion at multiple intermediate layers. For instance, CentralNet (Vielzeuf et al. 2018) and MMTM (Joze et al. 2020) join the latent representations at each layer and pass them as auxiliary information for deeper layers. Such methods achieve superior performances on several multimodal tasks including multimodal action recognition (Shahroudy et al. 2016) and gesture recognition (Zhang et al. 2018). However, it would largely increase the parameters of multimodal fusion models.

In recent years, there is an increased interest of introducing the attention mechanisms such as Transformer (Vaswani et al. 2017) to multimodal learning. The multimodal-BERT family (Chen et al. 2019; Li et al. 2019; Lu et al. 2019; Tan and Bansal 2019) is a typical approach for inter-modal fusion. Moreover, DFAF (Gao et al. 2019) shows that intramodal fusion could also be helpful. DFAF proposes a dynamic attention flow module to mix inter-modal and intramodal features together through the multi-head attention (Vaswani et al. 2017). Additional efforts are made to enhance multimodal fusion efficacy of attention mechanisms. For instance, AoANet (Huang et al. 2019) proposes the attention on attention (AoA) module, showing that adding an attention operation on top of another one could achieve better performance on image captioning task.

Recently, the NAS approaches are making an exciting progress for DNNs, and it shows a huge potential to introduce NAS to multimodal learning. One representative work is MFAS (Pérez-Rúa et al. 2019), which employs SMBO algorithm (Hutter, Hoos, and Leyton-Brown 2011) to search multimodal fusion strategies given the unimodal backbones. But as SMBO is a black-box optimization algorithm, every update step requires a bunch of DNNs to be trained, leading to the inefficiency of MFAS. Besides, MFAS only use concatenation and fully connected (FC) layers for unimodal feature fusion, and the stack of FC layers would be a heavy burden for computing. Further work like MMIF (Peng et al. 2020) and 3D-CDC (Yu et al. 2021) adopt the efficient DARTS algorithm (Liu, Simonyan, and Yang 2019) for architecture search but only support the search of unary operations on graph edges and use summation on every intermediate node for reduction. MMnas (Yu et al. 2020) allows searching the attention operations but the topological structure of the network is fixed during architecture search.

Different from these related works, our proposed BM-NAS supports to search both the unimodal feature selection strategy and the fusion strategy of multimodal DNNs. BM-NAS introduces a bilevel searching scheme. The upper level of BM-NAS supports both intra-modal and inter-modal feature selection. The lower level of BM-NAS searches the fusion operations within every intermediate step. Each step can flexibly form the summation, concatenation, multihead attention (Vaswani et al. 2017), attention on attention (Huang et al. 2019), or any other unexplored fusion mechanisms. BM-NAS is a generalized and efficient NAS framework for multimodal learning. In experiments we show that BM-NAS can be applied to various multimodal tasks regardless of the modalities or backbone models.

Methodology

In this work, we propose a generalized NAS framework, named Bilevel Multimodal NAS (BM-NAS), to search the architectures of multimodal fusion DNNs. More specifically, BM-NAS searches a Cell-by-Cell architecture in a bilevel fashion. The upper level architecture is a directed acyclic graph (DAG) of the input features and Cells. The lower level architecture is a DAG of inner step nodes within a Cell. Each inner step node is a bivariate operation drawn from a predefined pool. The bilevel searching scheme ensures that BM-NAS can be easily adapted to various multimodal learning tasks regardless of the types of modalities. In the following, we discuss the unimodal feature extraction, the upper and lower levels of BM-NAS, along with the architecture search algorithm and evaluation.

Unimodal Feature Extraction

By following previous multimodal fusion works, such as CentralNet (Vielzeuf et al. 2018), MFAS (Pérez-Rúa et al. 2019) and MMTM (Joze et al. 2020), we also employ the pretrained unimodal backbone models as the feature extractors. We use the outputs of their intermediate layers as raw features (or intermediate blocks if the model has a block-by-block structure like ResNeXt (Xie et al. 2017)).

Since the raw features vary in shapes, we reshape them by applying pooling, interpolation, and fully connected layers on spatial, temporal, and channel dimensions, successively. By doing so, we reshape all the raw features to the shape of (N, C, L), such that we can easily perform fusion operations between features of different modalities. Here N is the batch size, C is the embedding dimension or the number of channels, L is the sequence length.

Upper Level: Cells for Feature Selection

The upper level of BM-NAS searches the unimodal feature selection strategy and it consists of a group of Cells. Formally, suppose we have two modalities A and B, and two pretrained unimodal models for each modality. Let $\{A^{(i)}\}$ and $\{B^{(i)}\}$ indicate the modality features extracted by the backbone models. We formulate the upper level nodes in an ordered sequence $\mathcal{S}$, as

$$\mathcal{S} = [A^{(1)}, ..., A^{(N_A)}, B^{(1)}, ..., B^{(N_B)}, \text{Cell}^{(1)}, ..., \text{Cell}^{(N)}].$$

Under the setting of $\mathcal{S}$, both inter-modal fusion and intramodal fusion are considered in BM-NAS.

Feature Selection. By adopting the continuous relaxation in differentiable architecture search scheme (Liu, Simonyan, and Yang 2019), all predecessors of $\text{Cell}^{(i)}$ will be connected

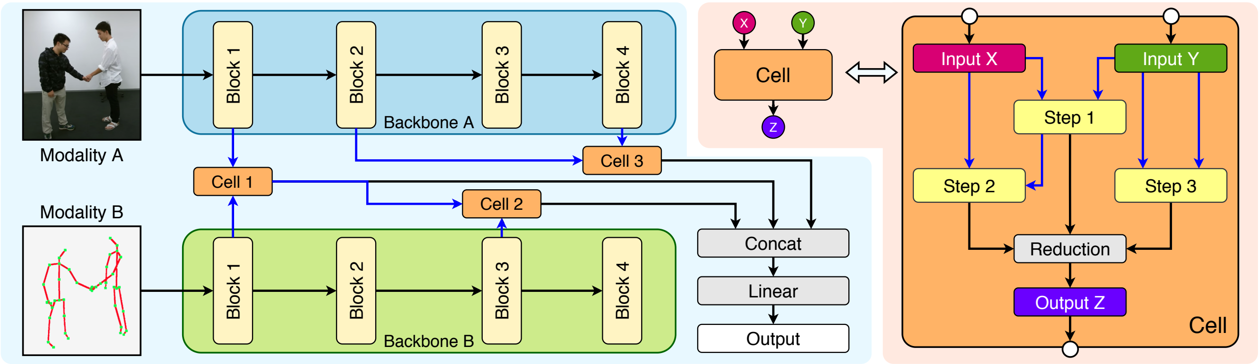

Figure 3: An example of a multimodal fusion network found by BM-NAS, which consists of a bilevel searching scheme, we denote searched edges in blue, and fixed edges in black. **Left**: The upper level BM-NAS. The input features are extracted by pretrained unimodal models. Each *Cell* accepts two inputs from its predecessors, *i.e.*, any unimodal feature or previous Cell. **Right**: The lower level BM-NAS. Within a Cell, each *Step* denotes a primitive operation selected from a predefined operation pool. The topologies of Cells and Steps are both searchable. The numbers of Cells and Steps are hyper-parameters such that BM-NAS can be adapted to a variety of multimodal tasks with different scales.

to Cell$^{(i)}$ through weighted edges at the searching stage. This directed complete graph between Cells is called the *hypernet*. For two upper level nodes $s^{(i)}, s^{(j)} \in \mathcal{S}$, let $\alpha^{(i,j)}$ denote the edge weight between $s^{(i)}$ and $s^{(j)}$. Each edge is a unary operation g selected from a function set $\mathcal{G}$ including

(1) $\mathrm{Identity}(x) = x$, *i.e.*, selecting an edge.

(2) $\mathrm{Zero}(x) = 0$, *i.e.*, discarding an edge.

Then, the mixed edge operation $\bar{g}^{(i,j)}$ on edge (i,j) is

$$\bar{g}^{(i,j)}(s) = \sum_{g\in\mathcal{G}} \frac{\exp(\alpha_g^{(i,j)})}{\sum\limits_{g'\in\mathcal{G}} \exp(\alpha_{g'}^{(i,j)})} g(s). \tag{1}$$

A Cell $s^{(j)}$ receives inputs from all its predecessors, as

$$s^{(j)} = \sum_{i<j} \bar{g}^{(i,j)}(s^{(i)}). \tag{2}$$

In evaluation stage, the network architecture is discretized that an input pair $(s^{(i)}, s^{(j)})$[1] will be selected for $s^{(k)}$ if

$$(i,j) = \underset{i<j<k,\ g\in\mathcal{G}}{\arg\max}\ (\alpha_g^{(i,k)} \cdot \alpha_g^{(j,k)}). \tag{3}$$

It is worth noting that, compared with searching the feature pairs directly, the Cell-by-Cell structure significantly reduces the complexity of the search space for unimodal feature selection. For an input pair from two feature sequences $[A^{(1)}, ..., A^{(N_A)}]$ and $[B^{(1)}, ..., B^{(N_B)}]$, the number of candidate choices is $2(N_A + N_B)$ under the Cell-by-Cell search setting. It is much smaller than $C^2_{N_A+N_B}$, the number of candidates under the pairwise search setting.

Lower Level: Multimodal Fusion Strategy

The lower level of BM-NAS searches the multimodal fusion strategy, *i.e.*, the inner structure of Cells. Specifically, a Cell is a DAG consisting of a set of inner step nodes. The inner step nodes are the primitive operations drawn from a predefined operation pool. We introduce our predefined operation pool in the following.

Primitive Operations. All the primitive operations take two tensor inputs x, y, and outputs a tensor z, where $x, y, z \in \mathbb{R}^{N\times C\times L}$.

(1) $\mathrm{Zero}(x, y)$: The Zero operation discards an inner step completely. It will be helpful when BM-NAS decides to use only a part of the inner steps.

$$\mathrm{Zero}(x, y) = 0. \tag{4}$$

(2) $\mathrm{Sum}(x, y)$: The DARTS (Liu, Simonyan, and Yang 2019) framework uses summation to combine two features as

$$\mathrm{Sum}(x, y) = x + y. \tag{5}$$

(3) $\mathrm{Attention}(x, y)$: We use the scaled dot-product attention (Vaswani et al. 2017). As a standard attention module usually takes three inputs namely query, key, and value, we let the query be x, the key and value be y, which is also known as the guided-attention (Yu et al. 2020).

$$\mathrm{Attention}(x, y) = \mathrm{Softmax}(\frac{xy^T}{\sqrt{C}} y). \tag{6}$$

(4) $\mathrm{LinearGLU}(x, y)$: A linear layer with the gated linear unit (GLU) (Dauphin et al. 2017). Let $W_1, W_2 \in \mathbb{R}^{C\times C}$ and $\odot$ be element-wise multiplication, then LinearGLU is

$$\begin{aligned} \mathrm{LinearGLU}(x, y) &= \mathrm{GLU}(xW_1, yW_2) \\ &= xW_1 \odot \mathrm{Sigmoid}(yW_2). \end{aligned} \tag{7}$$

(5) $\mathrm{ConcatFC}(x, y)$: ConcatFC stands for passing the concatenation of (x, y) to a fully connected (FC) layer with ReLU activation (Nair and Hinton 2010). The FC layer reduces the channel numbers from $2C$ to C. Let $W \in \mathbb{R}^{2C\times C}$, $b \in \mathbb{R}^C$, then ConcatFC is

$$\mathrm{ConcatFC}(x, y) = \mathrm{ReLU}(\mathrm{Concat}(x, y)W + b). \tag{8}$$

[1] We enforce the Cells to have different predecessors.

We elaborately choose these primitive operations such that they can be flexibly combined to form various feature fusion modules. In Fig. 2, we show that the search space of lower level BM-NAS accommodates many benchmark multimodal fusion strategies such as the summation used in DARTS (Liu, Simonyan, and Yang 2019), the ConcatFC used in MFAS (Pérez-Rúa et al. 2019), the multi-head attention used in Transformer (Vaswani et al. 2017), and the Attention on Attention used in AoANet (Huang et al. 2019). There also remains flexibility to discover other better fusion modules for specific multimodal learning tasks.

Fusion Strategy. In searching stage, the inner step set of $\text{Cell}^{(n)}$ is an ordered feature sequence $\mathcal{T}_n$,

$$\mathcal{T}_n = [x, y, \text{Step}^{(1)}, ..., \text{Step}^{(M)}]. \qquad (9)$$

An inner step node $t^{(i)}$ transforms two input nodes $t^{(j)}, t^{(k)}$ to its output through an average over the primitive operation pool $\mathcal{F}$, as

$$\overline{f}^{(i)}(t^{(j)}, t^{(k)}) = \sum_{f \in \mathcal{F}} \frac{\exp(\gamma_f^{(i)})}{\sum\limits_{f' \in \mathcal{F}} \exp(\gamma_{f'}^{(i)})} f(t^{(j)}, t^{(k)}), \qquad (10)$$

where γ is the weights of primitive operations. In the evaluation stage, the optimal operation of an inner step node is derived as,

$$f^{(i)} = \arg\max_{f \in \mathcal{F}} \gamma_f^{(i)}. \qquad (11)$$

The continuous relaxation of the edges with weights β between inner step nodes is similar to the upper level. For a simplicity, we omit the formulation in this paper. Note that unlike the upper level BM-NAS, the pairwise inputs in a Cell can be chosen repeatedly[2], so the inner steps can form structures like multi-head attention (Vaswani et al. 2017).

Architecture Search and Evaluation

Architecture Parameters. The function of the weights of primitive operations (β) and inner step nodes edges (γ) is shown in Fig. 4, β is used for feature selection within the cell, selecting two inputs for each inner step node. And γ is used for operation selection on each inner step node.

Search Algorithm. We introduced three variable α, β, γ as the architecture parameters. Algorithm 1 shows the searching process of BM-NAS, which follows DARTS (Liu, Simonyan, and Yang 2019) to optimize α, β, γ and model weights w, alternatively. In Algorithm 1, the model in searching stage is called *hypernet* since all the edges and nodes are mixed operations. The searched structure description of the fusion network is called *genotype*.

Implementation Details. In order to make the whole BM-NAS framework searchable and flexible, Cells/inner step nodes should have the same number of inputs and output, so they can be put together in arbitrary topological order. The two-input setting follows the benchmark NAS frameworks (DARTS (Liu, Simonyan, and Yang 2019)), MFAS (Pérez-Rúa et al. 2019), MMIF (Peng et al. 2020), *etc.*). They all have only two searchable inputs for each Cell/step node. Also, it requires no extra effort to let the Cells or step nodes support 3 or more inputs, by just adding ternary (or other arbitrary) operations into the primitive operation pool.

Algorithm 1: Bilevel Multimodal NAS (BM-NAS)

```
Result: The genotype of fusion networks.
Initialize architecture parameters $\alpha, \beta, \gamma$ and model
  parameters $w$;
Initialize genotype based on $\alpha, \beta, \gamma$, set
  genotype_best = genotype;
Construct hypernet based on genotype_best;
while $\mathcal{L}$ not converged do
    Update $\omega$ on training set;
    Update ($\alpha$, $\beta$, $\gamma$) on validation set;
    Derive upper level genotype based on $\alpha$, derive
      lower level genotype based on $\beta, \gamma$;
    Update hypernet based on genotype;
    if higher validation accuracy is reached then
        Update genotype_best using genotype;
    end
end
Return genotype_best;
```

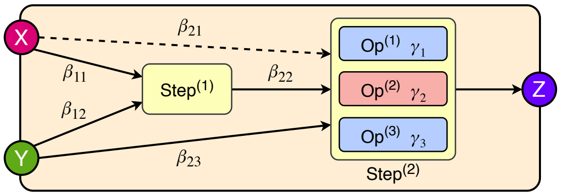

Figure 4: Architecture parameters β and γ of a *Cell*.

Evaluation. In architecture evaluation, we select the *genotype* with the best validation performance as the searched fusion network. Then we combine the training and validation sets together to train the unimodal backbones and the searched fusion network jointly.

Experiments

In this work we evaluate the BM-NAS on three multimodal tasks, including (1) the multi-label movie genre classification task on MM-IMDB dataset (Arevalo et al. 2017), (2) the multimodal action recognition task on NTU RGB-D dataset (Shahroudy et al. 2016), and (3) the multimodal gesture recognition task on EgoGesture dataset (Zhang et al. 2018). In the following, we discuss the experiments on the three tasks respectively. We perform computing efficiency analysis. We further evaluate the search strategies of the proposed BM-NAS framework.

In addition, we present the examples of these tasks, thorough discussion of hyper-parameter configurations, visualization of searched architectures and their performances during the searching stage (hypernets) and evaluation stage (final model) in our supplementary material.

[2] We don't enforce the step nodes to have different predecessors.

Method	Modality	F1-W(%)
Unimodal Methods		
Maxout MLP (ICML13)	Text	57.54
VGG Transfer (ICLR15)	Image	49.21
Multimodal 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 (ours)	Image + Text	**62.92 ± 0.03**

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-occurrence (IJCAI18)	Pose	85.24
Multimodal 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 (ours)	Video + Pose	**90.48 ± 0.24**

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

MM-IMDB Dataset

MM-IMDB dataset (Arevalo et al. 2017) is a multi-modal dataset collected from the Internet Movie Database, containing posters, plots, genres and other meta information of 25,959 movies. We conduct multi-label genre classification on MM-IMDB using posters (RGB images) and plots (text) as the input modalities. There are 27 non-mutually exclusive genres in total, including *Drama, Comedy, Romance*, etc. Since the number of samples in each class is highly imbalanced, we only use 23 genres for classification. The classes of *News, Adult, Talk-Show, Reality-TV* are discarded since they only count for 0.10% in total. We adopt the original split of the dataset where 15,552 movies are used for training, 2,608 for validation and 7,799 for testing.

For a fair comparison with other explicit multimodal fusion methods, we use the same backbone models. Specifically, we use Maxout MLP (Goodfellow, Warde-Farley, and Courville 2013) as the backbone of text modality and VGG Transfer (Simonyan and Zisserman 2015) as the backbone of RGB image modality. For BM-NAS, we adopt a setting of 2 fusion Cells and 1 step/Cell. For inner step representations, we set $C = 192, L = 16$.

Table 1 shows that BM-NAS achieves the best Weighted F1 score in comparison with the existing multimodal fusion methods. Notice that as the class distribution of MM-IMDB is highly imbalanced, Weighted F1 score is in fact a more reliable metric for measuring the performance of multi-label classification than other kinds of F1 score.

NTU RGB-D Dataset

The NTU RGB-D dataset (Shahroudy et al. 2016) is a large scale multimodal action recognition dataset, containing a total of 56,880 samples with 40 subjects, 80 view points, and 60 classes of daily activities. In this work we use the skeleton and RGB video modality for fusion experiments. We measure the performance of methods using cross-subject (CS) accuracy. We follow the dataset split of MFAS (Pérez-Rúa et al. 2019). In detail, we use subjects 1, 4, 8, 13, 15, 17, 19 for training, 2, 5, 9, 14 for validation, and the rest for test. There are 23760, 2519 and 16558 samples in the training, validation, and test dataset, respectively.

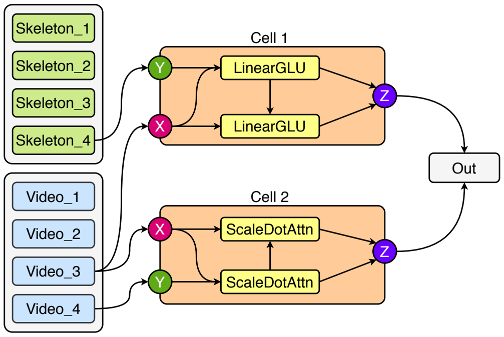

Figure 5: Best model found on NTU RGB-D dataset.

For a fair comparison, we use two CNN models, the Inflated ResNet-50 (Baradel et al. 2018) for video modality and Co-occurrence (Li et al. 2018) for skeleton modality as backbones, ensuring all the methods in our experiments share the same backbones. We test the performances of MFAS (Pérez-Rúa et al. 2019), MMTM (Joze et al. 2020), and the proposed BM-NAS using our data prepossessing pipeline, such that the performances of these methods are not the same as they were original reported. For BM-NAS, we use 2 fusion Cells and 2 Steps/Cell. For inner step representations we set $C = 128, L = 8$.

In Table 2, our method achieves an cross-subject accuracy of 90.48%, showing an state-of-the-art result on NTU RGB-D (Shahroudy et al. 2016) with video and pose modalities.

EgoGesture Dataset

The EgoGesture dataset (Zhang et al. 2018) is a large scale multimodal gesture recognition dataset, containing 24,161 gesture samples of 83 classes collected from 50 distinct subjects and 6 different scenes. We follow the original cross-subject split of EgoGesture dataset (Zhang et al. 2018). There are 14,416 samples for training, 4,768 for validation, and 4,977 for testing.

We use the ResNeXt-101 (Köpüklü et al. 2019) as the backbone on both RGB and depth video modality. As former works like CentralNet (Vielzeuf et al. 2018) and MFAS (Pérez-Rúa et al. 2019) did not perform experiments on this dataset, we compared our method with other unimodal and multimodal methods, especially MMTM (Joze et al. 2020), MTUT (Gupta et al. 2019) and 3D-CDC (Yu et al. 2021). Since we do not search for the backbone, we compared with 3D-CDC-NAS2, which also uses ResNeXt-101 as the backbones. For our BM-NAS, we use 2 fusion Cells and 3 steps/Cell, for inner step representations we set $C = 128, L = 8$.

Table 3 reports the experiment results on EgoGesture (Zhang et al. 2018). Comparing to 3D-CDC, which requires

Method	Modality	Acc(%)
Unimodal Methods		
VGG-16 + LSTM (NIPS14)	RGB	74.70
C3D + LSTM + RSTTM (ICCV15)	RGB	89.30
I3D (CVPR17)	RGB	90.33
ResNext-101 (FG19)	RGB	93.75
VGG-16 + LSTM (CVPR14)	Depth	77.70
C3D + LSTM + RSTTM (CVPR16)	Depth	90.60
I3D (CVPR17)	Depth	89.47
ResNeXt-101 (FG19)	Depth	94.03
Multimodal Methods		
VGG-16 + LSTM (CVPR17)	RGB + Depth	81.40
C3D + LSTM + RSTTM (CVPR19)	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 (ours)	RGB + Depth	**94.96 ± 0.07**

Table 3: Gesture recognition results on EgoGesture dataset. We use ResNext-101 as backbones for both RGB and depth modality for our BM-NAS method.

Method	Dataset	Parameters	Acc(%)
MMTM (CVPR20)	NTU	8.61 M	88.92
MFAS (CVPR19)	NTU	2.16 M	89.50
BM-NAS (ours)	NTU	**0.98 M**	**90.48**

Table 4: Model size and performance on NTU RGB-D.

Method	MM-IMDB	NTU
MFAS (CVPR19)	9.24	603.64
BM-NAS (ours)	**0.89**	**38.6**

Table 5: Search cost (GPU·hours) of generalized multimodal NAS methods.

3 groups of backbone models trained under different video frame rates (8, 16 and 32 FPS), our BM-NAS only requires the 32 FPS ones, and is generalized to all kinds of modalities. In general, BM-NAS achieves a state-of-the-art fusion performance, showing that BM-NAS is effective for enhancing gesture recognition performance on EgoGesture dataset.

Computing Efficiency

Model Size. Table 4 compares the model sizes of different multimodal fusion methods on NTU RGB-D (Shahroudy et al. 2016). All three methods share exactly the same unimodal backbones. Compared with the manually designed fusion model MMTM (Joze et al. 2020) and the fusion model searched by MFAS (Pérez-Rúa et al. 2019), our BM-NAS achieves better performance with fewer model parameters.

Search Cost. Table 5 compares the search cost of generalized multimodal NAS frameworks including MFAS and our BM-NAS. Thanks to the efficient differentiable architecture search framework (Liu, Simonyan, and Yang 2019), BM-NAS is at least 10x faster than MFAS when searching on MM-IMDB (Arevalo et al. 2017) and NTU RGB-D.

Ablation Study

In this section, we conduct ablation study to verify the effectiveness of the unimodal feature selection strategy and the multimodal fusion strategy, respectively.

Features	Dataset	Accuracy(%)
Random	NTU	86.35 ± 0.68
Late fusion	NTU	89.49 ± 0.15
Searched (MFAS)	NTU	89.50 ± 0.60
Searched (BM-NAS)	NTU	**90.48 ± 0.24**

Table 6: Ablation study for feature selection.

Fusion	Framework	Dataset	Acc (%)
Sum	DARTS (ICLR19)	NTU	87.64
ConcatFC	MFAS (CVPR19)	NTU	89.20
MHA	Transformer (NIPS17)	NTU	88.29
AoA	AoANet (ICCV19)	NTU	89.11
Searched	BM-NAS	NTU	**90.48**

Table 7: Ablation study for fusion strategy.

Unimodal Feature Selection. Table 6 compares different unimodal feature selection strategies on NTU RGB-D. We compare the best strategy found by BM-NAS against random selection, late fusion, and the best strategy found by MFAS. For all the random baselines, the inner structure of Cells are the same. We randomly selects the input features and the connections between Cells, and report the result averaged over 5 trials. For the late fusion baseline, we concatenate feature pair (*Video_4*,*Skeleton_4*) in Fig. 5. MFAS selects four feature pairs: (*Video_4*, *Skeleton_4*), (*Video_2*, *Skeleton_4*), (*Video_2*, *Skeleton_2*), and (*Video_4*, *Skeleton_4*). As shown in Table 6, the searched feature selection strategy is better than all baselines, demonstrating that a better unimodal feature selection strategy benefits the multimodal fusion performance. As shown in Table 6, the searched feature selection strategy is better than all baselines, demonstrating that a better unimodal feature selection strategy benefits the multimodal fusion performance.

Multimodal Fusion Strategy. Table 7 evaluates different multimodal fusion strategies on NTU RGB-D. All the strategies in Table 7 adopt the same feature selection strategy. We compare the best Cell structure found by BM-NAS against the summation used in DARTS (Liu, Simonyan, and Yang 2019), the ConcatFC used in MFAS (Pérez-Rúa et al. 2019), the multi-head attention (MHA) used in Transformer (Vaswani et al. 2017), and the attention on attention (AoA) used in AoANet (Huang et al. 2019). All these fusion strategies can be formed as certain combinations of our predefined primitive operations, as shown in Fig. 2. In Table 7, the fusion strategy derived by BM-NAS outperforms the baseline strategies, showing the effectiveness of searching fusion strategy for multimodal fusion models.

Conclusion

In this paper, we have presented a novel multimodal NAS framework BM-NAS to learn the architectures of multimodal fusion models via a bilevel searching scheme. To our best knowledge, BM-NAS is the first NAS framework that supports to search both the unimodal feature selection and the multimodal fusion strategies for multimodal DNNs. In experiments, we have demonstrated the effectiveness and efficiency of BM-NAS on various multimodal learning tasks.

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *ICML*, 173–182.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*, 469–478.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *ICML*, 933–941.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*, 6639–6648.

Goodfellow, I. J.; Warde-Farley, D.; and Courville, M. M. A. 2013. Bengio, Yoshua. Maxout networks. In *ICML*, 1319–1327.

Gupta, V.; Dwivedi, S. K.; Dabral, R.; and Jain, A. 2019. Progression Modelling for Online and Early Gesture Detection. In *3DV*, 289–297. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*, 4634–4643.

Hutter, F.; Hoos, H. H.; and Leyton-Brown, K. 2011. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, 507–523. Springer.

Jin, T.; Huang, S.; Chen, M.; Li, Y.; and Zhang, Z. 2020. SBAT: Video Captioning with Sparse Boundary-Aware Transformer. In *IJCAI*.

Jin, T.; Huang, S.; Li, Y.; and Zhang, Z. 2019. Low-Rank HOCA: Efficient High-Order Cross-Modal Attention for Video Captioning. In *EMNLP*, 2001–2011.

Joze, H. R. V.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: Multimodal Transfer Module for CNN Fusion. In *CVPR*, 13289–13299.

Kandasamy, K.; Neiswanger, W.; Schneider, J.; Poczos, B.; and Xing, E. P. 2018. Neural architecture search with bayesian optimisation and optimal transport. In *NeurIPS*, 2016–2025.

Köpüklü, O.; Gunduz, A.; Kose, N.; and Rigoll, G. 2019. Real-time hand gesture detection and classification using convolutional neural networks. In *FG*, 1–8. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*.

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*, 786–792.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.-J.; Fei-Fei, L.; Yuille, A.; Huang, J.; and Murphy, K. 2018a. Progressive neural architecture search. In *ECCV*, 19–34.

Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.-J.; Fei-Fei, L.; Yuille, A.; Huang, J.; and Murphy, K. 2018b. Progressive neural architecture search. In *ECCV*, 19–34.

Liu, H.; Simonyan, K.; and Yang, Y. 2019. Darts: Differentiable architecture search.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 13–23.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 289–297.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Natarajan, P.; Wu, S.; Vitaladevuni, S.; Zhuang, X.; Tsakalidis, S.; Park, U.; Prasad, R.; and Natarajan, P. 2012. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 1298–1305. IEEE.

Negrinho, R.; and Gordon, G. 2017. Deeparchitect: Automatically designing and training deep architectures. *arXiv preprint arXiv:1704.08792*.

Peng, Y.; Bi, L.; Fulham, M.; Feng, D.; and Kim, J. 2020. Multi-modality Information Fusion for Radiomics-Based Neural Architecture Search. In *MICCAI*, 763–771. Springer.

Pérez-Rúa, J.-M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. Mfas: Multimodal fusion architecture search. In *CVPR*, 6966–6975.

Pham, H.; Guan, M. Y.; Zoph, B.; Le, Q. V.; and Dean, J. 2018. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*.

Real, E.; Aggarwal, A.; Huang, Y.; and Le, Q. V. 2019. Regularized evolution for image classifier architecture search. In *AAAI*, volume 33, 4780–4789.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 1010–1019.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 568–576.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Teney, D.; Anderson, P.; He, X.; and Van Den Hengel, A. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, 4223–4232.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2018. Centralnet: a multilayer approach for multimodal fusion. In *ECCV*, 0–0.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*, 1492–1500.

Yan, K.; Ji, L.; Luo, H.; Zhou, M.; Duan, N.; and Ma, S. 2021. Control Image Captioning Spatially and Temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2014–2025.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*, 4651–4659.

Yu, Z.; Cui, Y.; Yu, J.; Wang, M.; Tao, D.; and Tian, Q. 2020. Deep Multimodal Neural Architecture Search. *arXiv preprint arXiv:2004.12070*.

Yu, Z.; Zhou, B.; Wan, J.; Wang, P.; Chen, H.; Liu, X.; Li, S. Z.; and Zhao, G. 2021. Searching multi-rate and multimodal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*.

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*, 20(5): 1038–1050.

Zhou, X.; Huang, S.; Li, B.; Li, Y.; Li, J.; and Zhang, Z. 2019. Text guided person image synthesis. In *CVPR*, 3663–3672.

Zoph, B.; and Le, Q. V. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Zoph, B.; and Le, Q. V. 2017. Neural architecture search with reinforcement learning. In *ICLR*.