

Harmonic-NAS: Hardware-Aware Multimodal Neural Architecture Search on Resource-constrained Devices

Mohamed Imed Eddine Ghebriout

Higher National School of Computer Science - ESI ex INI

IM_GHEBRIOUT@ESI.DZ

Halima Bouzidi [†]

Université Polytechnique Hauts-de-France, LAMIH/CNRS

HALIMA.BOUZIDI@UPHF.FR

Smail Niar

Hamza Ouarnoughi

Université Polytechnique Hauts-de-France, LAMIH/CNRS, INSA Hauts-de-France

SMAIL.NIAR@UPHF.FR

HAMZA.OUARNOUGHI@UPHF.FR

Editors: Berrin Yanikoğlu and Wray Buntine

Abstract

The recent surge of interest surrounding Multimodal Neural Networks (MM-NN) is attributed to their ability to effectively process and integrate multiscale information from diverse data sources. MM-NNs extract and fuse features from multiple modalities using adequate unimodal backbones and specific fusion networks. Although this helps strengthen the multimodal information representation, designing such networks is labor-intensive. It requires tuning the architectural parameters of the unimodal backbones, choosing the fusing point, and selecting the operations for fusion. Furthermore, multimodality AI is emerging as a cutting-edge option in Internet of Things (IoT) systems where inference latency and energy consumption are critical metrics in addition to accuracy. In this paper, we propose *Harmonic-NAS*¹, a framework for the joint optimization of unimodal backbones and multimodal fusion networks with hardware awareness on resource-constrained devices. *Harmonic-NAS* involves a two-tier optimization approach for the unimodal backbone architectures and fusion strategy and operators. By incorporating the hardware dimension into the optimization, evaluation results on various devices and multimodal datasets have demonstrated the superiority of *Harmonic-NAS* over state-of-the-art approaches achieving up to $\sim 10.9\%$ accuracy improvement, $\sim 1.91x$ latency reduction, and $\sim 2.14x$ energy efficiency gain.

Keywords: Multimodal Learning, Data Fusion, Neural Architecture Search, Edge Computing.

1. Introduction

Our environment is continuously providing us with a broad stream of sensory modalities. Selecting and processing these modalities allows us to take action, react appropriately, and get insights into our surroundings. The term *modality* is used to describe the several forms that sensory information might take (e.g., visual, textual, acoustic) Liang et al. (2021). This *multimodal* perception paradigm has been brought to the sphere of Artificial Intelligence (AI) to bridge further the gap between human brain functioning and neural networks Huang et al. (2021). Recently, Multimodal Neural Networks (MM-NN) have captivated a lot of interest within the deep learning community since they have proven to be more accurate than their unimodal counterparts in several tasks such

[†] Corresponding author.

1. <https://github.com/Mohamed-Imed-Eddine/Harmonic-NAS>

as action recognition [Molchanov et al. \(2016\)](#), image-video captioning [Pramanick et al. \(2021\)](#), sentiment analysis [Gong et al. \(2023\)](#), and healthcare [Soenksen et al. \(2022\)](#). However, multimodal neural networks are computation- and memory-demanding. Thus, their deployment on Edge and Tiny devices is constrained by the availability of hardware resources [Hou et al. \(2022\)](#).

Designing unimodal neural networks is still challenging as it requires tuning a broad set of architectural parameters [Elsken et al. \(2019\)](#). The design landscape becomes more complex for multimodal neural networks [Pérez-Rúa et al. \(2019\)](#) as it typically involves various backbones and fusion networks for unimodal and multimodal feature selection and extraction, respectively. Each unimodal backbone and fusion network is characterized by a specific set of architectural parameters and assigned a particular role in the multimodal learning scheme. Furthermore, incorporating the hardware dimension into the design process of multimodal networks limits their ability to fuse a large quantity of information [Rashid et al. \(2023\)](#).

Recently, Neural Architecture Search (NAS) [Elsken et al. \(2019\)](#) has emerged as a data-driven approach to automate the design of neural networks by searching for the optimal set of architectural parameters within a predefined search space. Typical NAS approaches adopt evolutionary algorithms [Jian et al. \(2023\)](#) or differentiable architecture search [Liu et al. \(2019\)](#) as a search strategy. With the emergence of Edge-AI, Hardware-aware NAS [Cai et al. \(2018\)](#) also added hardware efficiency (e.g., latency, energy, memory) as an optimization objective. The NAS paradigm has been leveraged for multimodal networks since MFAS [Pérez-Rúa et al. \(2019\)](#) and MM-NAS [Yu et al. \(2020\)](#) in which the fusion architecture is searchable for visual-textual modalities. BM-NAS [Yin et al. \(2022\)](#) provides a more general framework to jointly search for the fusion architecture and operators. However, related works have yet to investigate making the MM-NN fully searchable -through unimodal backbones and multimodal fusion networks- across different modalities, tasks, and datasets. Furthermore, the hardware dimension still needs to be included in existing multimodal-NAS frameworks to ease the deployment on resource-constrained devices.

1.1. Novel Contributions

In this paper, we present *Harmonic-NAS*, a novel Hardware-aware NAS framework for the design of Multimodal Neural Networks on resource-constrained Edge devices. Our proposed framework encompasses the following novelties and contributions:

1. *Harmonic-NAS* co-optimizes the design of unimodal backbones and fusion networks to learn an effective joint embedding of features from multiple modalities.
2. We make the MM-NN fully searchable through a hierarchical search space for **(i)** Unimodal backbones, built upon the once-for-all supernet [Cai et al. \(2019\)](#) and **(ii)** Multimodal fusion networks, built upon the differentiable search space of DARTS [Liu et al. \(2019\)](#).
3. To solve the bi-level design space exploration problem, *Harmonic-NAS* includes a two-tier optimization, where the first search stage is an evolutionary algorithm for unimodal backbone networks, whereas the second search stage is a differentiable NAS for fusion networks, with an integrated hardware-related loss function in both search stages.
4. We demonstrate the efficiency of *Harmonic-NAS* by conducting experiments on various multimodal datasets and Edge devices. Empirical results have seen up to $\sim 10.9\%$ accuracy improvement, $\sim 1.91\times$ latency reduction, and $\sim 2.14\times$ energy gain, stipulating further the importance of the hierarchical design optimization for multimodal NNs on Edge devices.

2. Related Work

2.1. Multimodal Neural Networks

The multimodality paradigm involves feature fusion from multiple modalities to learn a joint embedding of global information. Initially, fusion approaches operate on the extreme levels of feature abstraction within the neural network on early layers with low-level features and on the last layers with high-level features. Early fusion operates on an input level, whereas late fusion operates on an output level using aggregation operators such as averaging or voting. As modern unimodal backbones are deeper and larger with features extracted at many levels, intermediate fusion has been introduced to provide more flexibility in the fusion position by operating on the intermediate feature-map level [Vielzeuf et al. \(2018\)](#). However, one challenge arises in determining the fusion placement as dense fusion networks [Yu et al. \(2020\)](#) fail to scale when the unimodal backbone networks deepen, resulting in an exponential increase of the fusion parameters. Fused data and fusion operators define the joint embedding granularity and quality. Simple features can be fused using sum, concatenation, or tensor operations [Liu et al. \(2018b\)](#). Nevertheless, for complex multimodal tasks, sophisticated fusion networks such as Attention [Nagrani et al. \(2021\)](#), Graph Neural Networks [Cai et al. \(2022\)](#), and Mixture-of-experts [Mustafa et al. \(2022\)](#) are needed to effectively learn the complex interactions between modalities.

2.2. Neural Architecture Search

Neural Architecture Search (NAS) aims to automate the design exploration of neural networks [Elsken et al. \(2019\)](#). The NAS is typically viewed as a black-box optimization taking as input the search space and the optimization objective (e.g., accuracy). As a search strategy, existing NAS frameworks generally employ evolutionary [Bouzidi et al. \(2023\)](#); [Odema et al. \(2023\)](#); [Jian et al. \(2023\)](#), or differentiable [Liu et al. \(2019\)](#) search algorithms. However, the NAS process is labor intensive, requiring many training-validation trials on the explored neural networks. To alleviate this problem, progressive shrinking [Liu et al. \(2018a\)](#) and once-for-all (OFA) supernet [Cai et al. \(2019\)](#) have been proposed. The once-for-all scheme is widely adopted for one-shot NAS [Dong and Yang \(2019\)](#), which consists of training a supernet comprising all the NN candidates once via weight-sharing and reusing the pretrained supernet to sample NNs during the search. While traditional NAS paradigms assume a discrete encoding of the NN architecture, Differentiable NAS (DARTS) [Liu et al. \(2019\)](#) proposed a continuous relaxation of the NN encoding, allowing for gradient-based optimization. The NAS paradigm has been incorporated first by MFAS [Pérez-Rúa et al. \(2019\)](#) to serve multimodal NNs. However, MFAS operates on a priori fixed backbones and only uses concatenation as a fusion operator while searching for the fusion positions. MM-NAS [Yu et al. \(2020\)](#) then followed up by refashioning the MM-NN into an encoder-decoder scheme with fully searchable fusion operators. Still, the unimodal backbones in MM-NAS are highly specialized and lack scalability to other types of neural networks. MUFASA [Xu et al. \(2021\)](#) has first attempted to jointly optimize the unimodal backbones and the fusion network using an evolutionary NAS and a Transformer backbone on the MIMIC-CCS dataset [Johnson et al. \(2016\)](#). Nevertheless, MUFASA targets one small dataset with a relatively simple search space. Moreover, their one-stage evolutionary search is not scalable to sophisticated backbones, fusion operators, and multimodal tasks. The recent BM-NAS framework [Yin et al. \(2022\)](#) provides a more general and scalable fusion search using differentiable NAS [Liu et al. \(2019\)](#). Nevertheless, BM-NAS operates on fixed unimodal backbones, overlooking opportunities for further performance gains from optimizing

the unimodal feature extraction. Additionally, the hardware dimension is still missing in existing multimodal NAS approaches for further deployment on resource-constrained devices.

2.3. Hardware Acceleration for Multimodal Neural Network

Modern IoT systems (e.g., wearable devices) comprise sensors gathering data from multiple modalities (e.g. image, audio, text). Processing these data requires over-parameterized multimodal networks with high computation and memory demands. Commodity Tiny and Edge devices have limited resources, burdening further the deployment of MM-NNs [Rashid et al. \(2023\)](#). Thus, hardware-aware optimization is needed when designing MM-NNs. Hardware-aware NAS approaches [Cai et al. \(2018\)](#); [Wu et al. \(2019\)](#) have contemplated integrating the hardware dimension into the NAS process for unimodal neural networks. Nonetheless, the multimodal case is still understudied. While few attempts have been made towards the optimization of MM-NN on Tiny and Edge devices by integrating latency as objective [Liu et al. \(2021\)](#), exploiting computation parallelism [Zhang et al. \(2022\)](#) and mixed-precision quantization [Rashid et al. \(2023\)](#), a holistic design exploration framework for $\text{HW} \times \text{MM-NN}$ is lacking in existing works.

Table 1: Comparison between existing works on Multimodal-NAS and ours.

Multimodal-NAS work	MFAS	MM-NAS	MUFASA	MMTM	BM-NAS	Harmonic-NAS (ours)
Unimodal backbone search			✓			✓
Unimodal feature selection	✓		✓	✓	✓	✓
Fusion micro-arch. search	✓	✓	✓	✓	✓	✓
Fusion macro-arch. search			✓			✓
Multimodal design flexibility					✓	✓
Multimodal tasks scalability		✓		✓	✓	✓
Hardware awareness						✓

Previous multimodal-NAS approaches in MFAS [Pérez-Rúa et al. \(2019\)](#), MM-NAS [Yu et al. \(2020\)](#), MUFASA [Xu et al. \(2021\)](#), MMTM [Joze et al. \(2020\)](#), and BM-NAS [Yin et al. \(2022\)](#) mostly rely on a priori fixed pretrained backbones as unimodal feature extractors and search only for the fusion network. However, this approach incurs the following drawbacks: **(i)** The a priori fixed unimodal backbones are not initially designed to serve multimodal networks. **(ii)** The Multimodal fusion performance depends highly on the quality of the extracted unimodal features. **(iii)** The overlooked performance and efficiency gains that can be obtained from specializing the multimodal fusion to less hardware-demanding backbones. Thus, *Harmonic-NAS* aims to make the MM-NN fully searchable by hierarchizing the architectural parameters related to the unimodal backbones and multimodal fusion networks design and optimizing them at once through a two-tier optimization strategy. *Harmonic-NAS* also includes the hardware dimension as an optimization objective to ease the deployment of MM-NNs on Edge devices that characterize modern IoT systems. In Table 1, we highlight the key differences between related works on multimodal-NAS and ours.

3. Methodology

In this paper, we propose *Harmonic-NAS*, a novel framework that aims to make the design of MM-NN fully searchable. As illustrated in Figure 1, *Harmonic-NAS* comprises a two-tier optimization stages: **(i)** The *first-stage* searches for optimal unimodal backbones for each *modality*. This involves exploring a search space of modality-specific backbone architectures and evaluating their

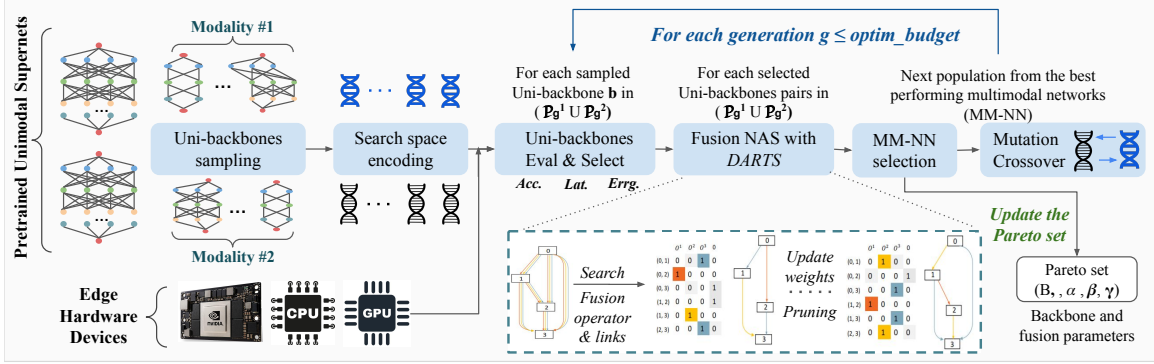


Figure 1: An overview of *Harmonic-NAS* for multimodal neural architecture search.

performance on the target task and their hardware efficiency on the Edge device. **(ii)** Within this process, the most promising unimodal backbones are selected for fusion network optimization in the *second-stage* to derive multimodal networks. The aim of the *second-stage* is to find a fusion network capable of effectively learning a joint embedding of features, thereby fully leveraging data’s intrinsic multimodality. *Harmonic-NAS* addresses this challenge by leveraging differentiable NAS (DARTS) to adapt the fusion network to the selected unimodal backbones. The two search stages are executed iteratively until reaching a final optimization budget (e.g., evolutionary generations).

3.1. First-stage: Unimodal Backbone Search

3.1.1. UNIMODAL BACKBONE DESIGN AND TRAINING

Given the heterogeneity of modalities, tasks, and hardware devices, a meticulous design of unimodal backbones is essential to achieve the optimal performance-efficiency tradeoff. To achieve this, *Harmonic-NAS* adopts a once-for-all approach by designing a *Supernet* comprising diverse neural architecture configurations. This will save valuable time for the unimodal backbones search step and facilitate the discovery of more efficient and task-specific NN architectures.

① **Supernet Design Specifications:** A modality-specific supernet \mathcal{S}_i is defined as a hyper-network of subnets sharing the same macro-architecture (i.e., neural blocks) and weights as depicted in Figure 2-(a). A unimodal backbone $\mathcal{B}_i(\cdot)$ (i.e., subnet from \mathcal{S}_i) for the i^{th} modality is represented as a succession of m neural blocks each comprising a sequence of layers \mathcal{L} as follows:

$$\mathcal{B}_i(\cdot) = \mathcal{B}_i^m \circ \mathcal{B}_i^{m-1} \circ \dots \circ \mathcal{B}_i^1 \quad | \quad \mathcal{B}_i^j = \mathcal{L}^{d_j} \circ \mathcal{L}^{d_j-1} \circ \dots \circ \mathcal{L}^1 \quad | \quad \text{for each } \mathcal{B}_i \in \mathcal{S}_i \quad (1)$$

Each block \mathcal{B}_i^j is characterized by a unique micro-architecture parameterized by a dynamic depth d_j , kernel sizes k_j , and channel expand ratio e_j . As we target deploying MM-NN on resource-constrained devices, we leverage the same search space based on the MobileNet-v3 baseline in Cai et al. (2019). We note that the backbone search space for the i^{th} modality is designated as \mathcal{S}_i .

② **Supernet Training:** Training a *Supernet* can be a challenging task. It differs from training a single NN since we jointly optimize the shared weights of all the subnets as follows:

$$\min_{\mathbb{W}} \sum_{\text{subnet}_i} \mathcal{L}_{\text{test}}(K(\mathbb{W}, \text{subnet}_i), \mathcal{Y}) \quad (2)$$

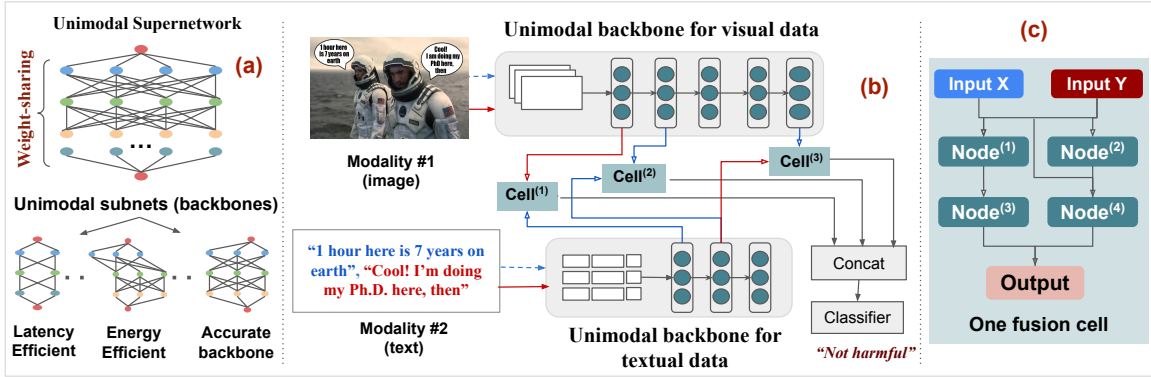


Figure 2: (a) An illustration of the unimodal supernet with subnets sharing weights and neural macro-architecture. (b) A high-level overview of the multimodal NN with modality-specific backbones and fusion network comprises multiple *cells*. (c) A detailed view of one fusion *cell* with multiple nodes, each assigned a particular fusion operator.

where \mathbb{W} denotes the shared weights from the supernet for the architectural configuration $subnet_i$ obtained by a sampling method denoted K , and \mathcal{Y} is the ground truths labels. The aim is to optimize \mathbb{W} for each sampled subnet while minimizing the cost of the independent training of subnets via the weight-sharing and adequate sampling scheme that maintains fairness across subnets.

We train our unimodal supernets using *knowledge distillation* from the largest subnet (i.e., teacher) to guide the training of smaller subnets (i.e., students) Wang et al. (2021). By distilling the knowledge from the largest subnet, we aim to transfer its learned unimodal feature representations to smaller subnets. We also employ the sandwich rule Yu and Huang (2019) to sample subnets at each epoch. This rule involves sampling two types of subnets – those following a random distribution Γ over the search space \mathcal{S}_i and those belonging to a predefined set \mathcal{J} from \mathcal{S}_i . Alternatively, we define the overall set of unimodal subnets evaluated at each epoch as follows:

$$\{\mathcal{B}_{im} | \mathcal{B}_{im} \sim \Gamma(\mathcal{S}_i)\}_{m=1}^{|\mathcal{S}_i|} \cup \{\mathcal{B}_{in} | \mathcal{B}_{in} \in \mathcal{J}\}_{n=1}^{|\mathcal{S}_i|} \quad (3)$$

Therefore, by considering our selection method, the training process of the unimodal supernet aims to optimize the shared weights \mathbb{W} by choosing a uniform distribution of subnets Γ in addition to the max-subnet and the min-subnet in \mathcal{J} . Our supernet training loss is depicted in equation 4.

$$\arg \min_{\mathbb{W}} \left(\sum_{\mathcal{B}_{im} \in \Gamma(\mathcal{S}_i)} \mathcal{L}(\mathcal{B}_{im}, \mathbb{W}, \mathcal{Y}) + \sum_{\mathcal{B}_{in} \in \mathcal{J}} \mathcal{L}(\mathcal{B}_{in}, \mathbb{W}, \mathcal{Y}) \right) \quad (4)$$

3.1.2. EVOLUTIONARY SEARCH STRATEGY

Once the unimodal supernets are fully trained, *Harmonic-NAS* employs an evolutionary search strategy at the *first-stage* to explore the design spaces of backbones. The evolutionary search is set to be run for a specific number of generations. For each generation, it creates populations $\mathcal{P}_g^{\mathcal{S}_i}$ of backbones - for each modality - from which the multimodal networks will be procured. A value-encoding step is also used to create discrete vectors characterizing the architectural parameters of the neural blocks (i.e., depth, kernel size, and channel expand ratio) of the sampled backbones.

Afterward, each backbone undergoes an evaluation to assess its unimodal performance on the validation set and its hardware metrics of the inference (i.e., latency and energy consumption). The hardware metrics are directly computed using device-specific lookup tables (LUTs). For each backbone $\mathcal{B}_i \in \mathcal{P}_g^{S_i}$, the evaluation function of the unimodal performance is defined as follows:

$$\mathcal{U}(\mathcal{B}_i) = Eval_score(Acc(\mathcal{B}_i), Lat(\mathcal{B}_i), Enrg(\mathcal{B}_i)) \quad (5)$$

To compute *Eval_score* in the underlying multi-objective context, *Harmonic-NAS* incorporates a Pareto ranking using the non-dominated sorting algorithm on the unimodal performance metrics (i.e., accuracy, latency, and energy) and a crowding distance that measures the diversity of backbones in the objective space [Deb et al. \(2000\)](#). The evaluation score is then used to identify a subset of the top-performing unimodal backbones $\mathcal{P}'_g^{S_i} \subset \mathcal{P}_g^{S_i}$ on which the second search stage for fusion network will be performed to derive optimal multimodal networks.

After the completion of the fusion search on $\mathcal{P}'_g^{S_i}$, *Harmonic-NAS* proceeds, through the evolutionary search engine, to the creation of the next population $\mathcal{P}_{g+1}^{S_i}$ of unimodal backbones using *mutation* and *crossover*. Within this process, a second selection criterion is set only to pick backbones achieving the best performance in their multimodal variant (i.e., with fusion network) as elite solutions for *mutation* and *crossover*. A uniform mutation is employed on the neural block level of backbones by sampling new depth, kernel size, and channel expand ratio under a probability threshold of 0.4. The crossover is applied by randomly picking two unimodal backbones -for the same modality- and swapping their neural blocks under a probability threshold of 0.8.

3.2. Second-stage: Multimodal Fusion Search

3.2.1. FUSION NETWORK SEARCH SPACE

In light of the scalability of *Harmonic-NAS* to the multimodal task complexity and diversity of backbone architectures, it's crucial to define a generalized search space with all the possible options for the fusion macro-architecture (i.e., number of fusion cells and nodes) and micro-architecture (i.e., fusion positions and operators). To this end, we adopt a cell-based fusion search space as introduced in [Yin et al. \(2022\)](#) with a parameterized fusion macro-architecture. As depicted in Figure 2-(b), the fusion network is built upon *fusion cells*, wherein each cell selects unimodal features to be fused. These unimodal features can be chosen from the output of intermediate blocks of the backbones or the outputs of previous fusion cells. A fusion cell comprises multiple fusion nodes, as shown in Figure 2-(c), each performing a specific fusion operator on the selected features.

① **Unimodal Feature Selection:** Assuming the availability of m modality each processed by a specific backbone B_1, \dots, B_m , respectively. A unimodal feature extracted by the j th neural block of the i th backbone is denoted as $B_i^{(j)}$. The unimodal feature selection procedure consists of choosing for each fusion $Cell^{(p)}$, two input features from the unimodal features set \mathcal{F}_1 as given by (6), enabling both inter and intra-modality fusion:

$$\mathcal{F}_1 = [B_1^{(1)}, \dots, B_1^{(N_{B_1})}, B_m^{(1)}, \dots, B_m^{(N_{B_m})}, Cell^{(1)}, \dots, Cell^{(p-1)}] \quad (6)$$

By performing a continuous relaxation on our search space, each fusion $Cell^{(p)}$ receives a weighted sum of the \mathcal{F}_1 elements with their corresponding probabilities (α) of being selected or not. These probabilities are updated during the training phase of the fusion network using DARTS. Then,

each fusion $\text{Cell}^{(p)}$ operates on the received weighted sum of input features as follows:

$$\text{Cell}^{(p)} = \sum_{k=0}^{|\mathcal{F}_1|} \bar{a}^{(k,p)}(\mathcal{F}_1[k]) \quad , \quad \bar{a}^{(k,p)}(s) = \sum_{a \in \mathcal{A}} \frac{\exp(\alpha_a^{(k,p)})}{\sum_{a' \in \mathcal{A}} \exp(\alpha_{a'}^{(k,p)})} a(s) \quad (7)$$

Here \mathcal{A} represents a set of two functions: Identity ($o(x) = x$) and Zero ($o(x) = 0$). At the evaluation phase, only the input features (X, Y) depicting the highest probabilities will be chosen as unimodal features as follows :

$$(X, Y) = \arg \max_{p < r < k, a \in \mathcal{A}} (\alpha_a^{(p,k)} \cdot \alpha_a^{(r,k)}) \quad (8)$$

② **Multimodal Feature Selection and Fusion Operators:** At this stage, we investigate the design of the inner structure of one fusion $\text{Cell}^{(p)}$. A fusion cell constitutes \mathcal{D} fusion nodes, each assigned a particular fusion operator from our predefined fusion operators set \mathcal{FP} (See Table 2). A fusion Node^(d) operates on two inputs – In the case of the first fusion node, the inputs are directly $\text{Cell}^{(p)}$'s inputs. However, for subsequent fusion nodes, the inputs can also be the outputs of previous fusion nodes. More formally, the inputs of a fusion Node^(d), are selected from the multimodal features set \mathcal{F}_2 defined as follows:

$$\mathcal{F}_2 = [X, Y, \text{Node}^{(1)}, \dots, \text{Node}^{(d-1)}].$$

where (X, Y) are the inputs of the current fusion $\text{Cell}^{(p)}$ while $\text{Node}^{(1, \dots, d-1)}$ denotes the output of the underlying fusion node within $\text{Cell}^{(p)}$. Similarly to the unimodal feature selection mechanism for the fusion cell, fusion nodes follow the same strategy to select their input features (x, y) , as shown in equation 7 expect that here we use the β weights instead of α and select inputs from \mathcal{F}_2 .

Another layer of complexity is added by searching further which fusion operator to be used at each fusion node. This is done by assuming (x, y) as Node^(d) inputs and applying a continuous relaxation over the fusion operators set \mathcal{FP} as follows:

$$\bar{f}^{(d)}(x, y) = \sum_{f \in \mathcal{FP}} \frac{\exp(\gamma_f^{(d)})}{\sum_{f' \in \mathcal{FP}} \exp(\gamma_{f'}^{(d)})} f(x, y) \quad (9)$$

where γ is the weight matrix that sets a priority score for each fusion operator to be selected for each node. At the evaluation phase, the following criterion is used to select the best fusion operator from the fusion set \mathcal{FP} :

$$f^{(d)} = \arg \max_{f \in \mathcal{FP}} \gamma_f^{(d)} \quad (10)$$

To further ensure the multimodal fusion effectiveness across various designs of backbones, we consider hardware efficiency as a criterion when defining \mathcal{FP} . We select fusion operators that minimize the hardware burden while ensuring effective multimodal fusion. Table 2 provides details on the employed fusion operators with a brief explanation of each operator:

1. Sum: In the context of multimodal fusion, there are instances where cross-modality interactions exhibit an additive nature. In scenarios where the modalities are relatively independent, aggregating their representations through summation offers a straightforward means of capturing joint information that incorporates each modality's distinctive strengths.

Table 2: The set \mathcal{FP} of the employed fusion operators with their respective equations.

Fusion Operator	Mathematical Formula
Sum	$\text{Sum}(X, Y) = X + Y$
Attention	$\text{ScaleDotAttn}(X, Y) = \text{Softmax}\left(\frac{XY^T}{\sqrt{C}}\right) \cdot Y$
LinearGLU	$\text{LinearGLU}(X, Y) = \text{GLU}(XW_1, YW_2) = XW_1 \odot \text{Sigmoid}(YW_2)$
ConcatFC	$\text{ConcatFC}(X, Y) = \text{ReLU}(\text{Concat}(X, Y) \cdot W + b)$
Squeeze – Excitation	$\text{SE}(X, Y) = E_X \odot Y \mid E_X = \sigma(S_X \cdot W + b) \mid S_X = \frac{1}{L} \sum_{i=1}^L X(B, C, i)$
ConcatMish	$\text{ConcatMish}(X, Y) = \text{Mish}(\text{Cat}(XW_1, YW_2)) \mid \text{Mish}(Z) = \tanh(\log(1 + \exp(X))) \odot \text{Sigmoid}(YW_2)$

2. **ScaleDotAttn**: Motivated by the promising outcomes of the Attention mechanism in modeling cross-modality interactions [Nagrani et al. \(2021\)](#), we incorporate the scaled dot-product attention as a fusion operator to investigate its efficiency in addressing multimodal tasks.
3. **LinearGLU**: Through performing element-wise multiplication \odot on the linearly transformed modality X and the sigmoid-gated of modality Y , this operator allows modality Y to determine the contribution of each element from the modality X .
4. **ConcatFC**: Here the unimodal features are concatenated on the channel dimension C . Then a linear transformation is applied and followed by the *ReLU* activation.
5. **Squeeze – Excitation**: The utilization of the Squeeze-and-Excitation module [Hu et al. \(2018\)](#) for channel-wise recalibration has demonstrated its effectiveness when applied on various blocks of the unimodal convolutional neural network (CNN). We further extend its applicability to the multimodal case and include it as a fusion operator.
6. **CatConvMish** : This operator provides a fusion mechanism that captures both linear and non-linear interactions between the modalities through a series of primitive operations.

3.2.2. DIFFERENTIABLE SEARCH STRATEGY

In section 3.2.1, we defined the search space for the fusion network. This search space also includes parameters related to the fusion macro-architecture, such as \mathcal{C} , the number of fusion cells, and \mathcal{D} , the number of fusion nodes within each cell. We note that these parameters are also searchable within the first evolutionary search stage, ensuring a diverse range of fusion macro-architectures. While in the *second-stage*, we search for the fusion micro-architecture using DARTS for a priori sampled combination of \mathcal{C} and \mathcal{D} in the *first-stage* of *Harmonic-NAS*.

During the *second-stage*, the weights α , β , and γ are jointly updated. This involves the use of gradient-based optimization, which allows for the exploration of various fusion micro-architecture configurations by training a hypernetwork with the following loss function \mathcal{L}_{fusion} :

$$\min_{\alpha, \beta, \gamma} \mathcal{L}_{fusion}(\mathcal{H}, Z, \alpha, \beta, \gamma, \mathcal{Y})$$

$$\mathcal{L}_{fusion}(\mathcal{H}, Z, \alpha, \beta, \gamma, \mathcal{Y}) = [\mathcal{L}_{task}(\mathcal{H}, Z, \alpha, \beta, \gamma, \mathcal{Y})]^a + [\mathcal{L}_{Lat}(\mathcal{H}, \gamma)]^b + [\mathcal{L}_{Enrg}(\mathcal{H}, \gamma)]^c$$

where \mathcal{H} is the initial fusion hypernet that represents all the possible configurations for DARTS, γ are the weights that control the fusion operator selection within the fusion cells, Z denotes the multimodal input data, and \mathcal{Y} are the ground truth labels. The exponents a , b , and c serve as control knobs for the importance of each performance metric in the overall loss function \mathcal{L}_{fusion} . By

adjusting these exponents, we can emphasize particular objectives such as task performance, latency, or energy consumption during the search process of the fusion network. The task loss \mathcal{L}_{task} varies depending on the target task (e.g., cross-entropy, binary cross-entropy). The hardware loss $\mathcal{L}_{(Lat||Enrg)}$ involves hardware-specific metrics – It takes as input the architectural specification of the fusion network and a lookup table (LUT_{device}) for latency and energy measurements of the fusion operators on the target device. The hardware loss is then computed as follows:

$$\mathcal{L}_{(Lat||Enrg)}(\mathcal{H}, \gamma, LUT_{device}) = \sum_{p=1}^{\mathcal{C}} \sum_{d=1}^{\mathcal{D}} \gamma_p[d] \cdot LUT_{device}(\mathcal{FP}, (Lat||Enrg)) \quad (11)$$

We also note that we first apply the *softmax* operation on the γ weights so that $\gamma_p[d]$ represents the probabilities of each fusion operator to be selected for the d^{th} node within the p^{th} cell, where LUT_{device} contains hardware measurements for each fusion operator within \mathcal{FP} .

4. Experiments

4.1. Experimental Setup

4.1.1. MULTIMODAL TASKS AND DATASETS

To evaluate the effectiveness of *Harmonic-NAS* in designing efficient multimodal networks, we conduct experiments on various multimodal datasets as listed in Table 3.

Table 3: Multimodal Datasets and tasks used by *Harmonic-NAS*

Dataset	Modalities	Samples (train; val; test)	Task
AV-MNIST	Image, Audio	{55000; 10000; 5000}	Digit classification
MM-IMDB	Image, Text	{15552; 2608; 7799}	Movie genres classification
HARM-P	Image, Text	{3020; 177; 355}	Harmful US-Politics-related memes detection

- ① **AV-MNIST:** The audio-visual MNIST for hand handwritten digits classification. It includes two modalities: image samples of handwritten digits and audio samples of spoken digits.
- ② **MM-IMDB:** The multi-modal IMDB dataset for multi-label classification of movie genres using movie titles and metadata as textual modality as well as movie posters as a visual modality.
- ③ **Harmful Memes:** We use the *Harm-P* dataset [Pramanick et al. \(2021\)](#) for detecting harmful memes related to United States politics. The dataset contains memes images collected from various social media platforms and text extracted from the meme image using Google’s OCR Vision API².

4.1.2. UNIMODAL BACKBONES SETTINGS

We build our unimodal supernet for image and audio processing upon the once-for-all framework [Cai et al. \(2019\)](#). We further adjust their original search space by reducing the number of neural blocks to 3 and 5 for the *AV-MNIST* and *Memes-P* datasets, respectively. For each block: The depth is chosen from $\{2, 3, 4\}$, the width expansion ratio for each layer within the block is selected from $\{3, 4, 6\}$, and the kernel size is picked from $\{3, 5, 7\}$. Overall, the search space complexity ranges between $\mathcal{O}(2 \times 10^5)$ and $\mathcal{O}(2 \times 10^7)$. For textual modality, we use a Maxout network to process the text embedding on MM-IMDB dataset and HARM-P. To train our backbones, we use Adam as an optimizer with weight decay of $1e^{-4}$ and a cosine annealing as a learning rate scheduler with a base learning rate of $1e^{-3}$. An early-stopping is used to determine the number of training epochs.

2. cloud.google.com/vision/docs/ocr

4.1.3. HARDWARE EXPERIMENTAL SETTINGS

The multimodal workloads are deployed using the Pytorch 1.12 framework running on top of CUDA 11.4 and cuDNN 8.3.2. We target the following Edge devices provided by NVIDIA: **(i)** Jetson AGX Xavier equipped with an NVIDIA Carmel Arm-64bit CPU and a high-performance Volta GPU of 512 GPU cores and 64 Tensor cores. **(ii)** Jetson TX2 composed of an NVIDIA Denver 64Bit and ARM-A57 CPU cores along a high-performance Pascal GPU of 256 GPU cores.

4.1.4. EVOLUTIONARY AND DIFFERENTIABLE SEARCH SETTINGS

In our optimization process, we run the evolutionary search for 30 generations, each with a population size of 128 individuals. The top quartile of the promising backbones is selected for the fusion search to derive multimodal networks. Then, half of the elite multimodal networks are chosen for mutation and crossover on their unimodal backbones to generate the next population of the evolutionary search. The mutation and crossover probabilities are set to 0.4 and 0.8, respectively. We run the differentiable fusion search using DARTS for 25 epochs using Adam as an optimizer and a cosine-annealing learning rate scheduler. The learning rate ranged from e^{-6} to e^{-4} .

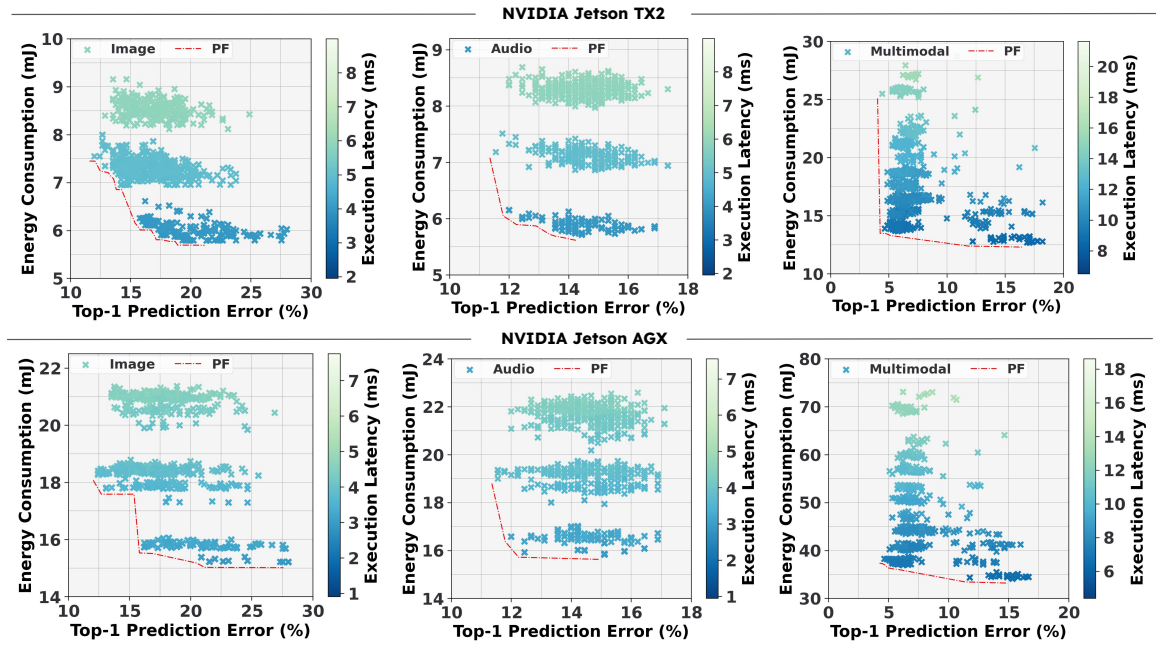


Figure 3: The first two columns (left to right) show the explored unimodal backbones in the first search stage for image and audio modalities, respectively. The last column shows the explored multimodal networks in the second search stage. The first and second rows report results on the NVIDIA TX2 and AGX devices, respectively. The red Pareto Front (PF) highlights the models with the best tradeoff between performance metrics.

4.2. Experimental Results

4.2.1. TWO-TIER OPTIMIZATION RESULTS ANALYSIS

To showcase the efficiency of the two-tier optimization of *Harmonic-NAS*, we report the search results for the AV-MNIST dataset on two hardware devices in Figure 3. The two first columns depict the explored unimodal backbones in the first search stage for image and audio, respectively, whereas the last column shows the explored multimodal networks in the second search stage. From the reported result, the explored unimodal backbones in the first search stage exhibit high variation in TOP-1 error ranges from $\sim 12\%$ to $\sim 27\%$ for image modality and from $\sim 12\%$ to $\sim 18\%$ for audio modality. By searching for optimal fusion networks for the two modalities in the second search stage, as shown in the third column, the TOP-1 error ranges from $\sim 4\%$ to $\sim 17\%$, improving further the performance of the unimodal backbones by up to **63%** error decrease. We can also notice the search intensification in the region that provides low TOP-1 errors in the multimodal case where more than $\sim 60\%$ of explored networks exhibit a TOP-1 error less than $\sim 8\%$. On the hardware efficiency, latency and energy values are generally doubled for the multimodal case as backbones and fusion are executed in a sequential pipeline manner. However, by integrating the hardware loss, *Harmonic-NAS* was able to identify a diverse set of optimal solutions in the Pareto front with a good compromise between prediction TOP-1 error, latency, and energy consumption.

4.2.2. PARETO OPTIMAL MULTIMODAL MODELS ANALYSIS

In this section, we provide an in-depth analysis of the Pareto optimal backbone and multimodal models obtained by *Harmonic-NAS* and compare them against SoTA multimodal-NAS approaches on different Edge devices. In the following, we analyze the results of each multimodal dataset.

① **The Audio-Visual MNIST:** Table 4 reports the obtained performance metrics (Top-1 Accuracy, latency, and energy) of the Pareto optimal backbones and multimodal models found by *Harmonic-NAS* compared against SoTA counterparts. In the rest of the paper, for abbreviation, latency and energy values on the Jetson AGX and TX2 are preceded by letters 'A' and 'T', respectively. Firstly, we noticed that the use of LeNet-based backbones in SoTA models such as [Wu and Goodman \(2018\)](#) and [Pérez-Rúa et al. \(2019\)](#) limits the unimodal feature extraction capability even if it's

Table 4: Performance evaluation on AV-MNIST dataset

SoTA work	Modality	Acc(%)	Latency(ms)	Energy(mJ)
Unimodal Backbones				
MVAE Wu and Goodman (2018)	Image	65.10	A:1.35,T:2.62	A:4.82,T:2.71
MFAS Pérez-Rúa et al. (2019)	Image	74.52	A:1.03,T:1.93	A:3.78,T:2.11
Harmonic-NAS (T:TX2)	Image	88.00	4.97	7.54
Harmonic-NAS (A:AGX)	Image	87.55	3.91	18.26
MVAE Wu and Goodman (2018)	Audio	42.00	A:1.92,T:4.12	A:9.73,T:11.71
MFAS Pérez-Rúa et al. (2019)	Audio	66.06	A:1.78,T:2.92	A:6.98,T:5.50
Harmonic-NAS (T:TX2)	Audio	88.44	4.97	7.18
Harmonic-NAS (A:AGX)	Audio	88.44	3.74	18.99
Multimodal Neural Networks				
MVAE Wu and Goodman (2018)	Image + Audio	72.30	A:4.40,T:8.53	A:15.64,T:15.67
MFAS Pérez-Rúa et al. (2019)	Image + Audio	88.38	A:4.37,T:6.41	A:10.76,T:10.67
BM-NAS Yin et al. (2022)	Image + Audio	91.11	A:3.26,T:5.41	A:14.17,T:8.60
		92.88	8.96	13.93
Harmonic-NAS (TX2)	Image + Audio	95.55	14.41	25.49
		95.33	9.11	13.88
		94.22	6.95	37.17
Harmonic-NAS (AGX)	Image + Audio	95.33	7.26	38.24
		95.11	7.19	38.14

boosted with a powerful fusion architecture search strategy as in BM-NAS [Yin et al. \(2022\)](#). Thus, the low-performing unimodal backbones limit the multimodal performances. However, by optimizing the unimodal backbones, an accuracy improvement of up to $\sim 4.44\%$ has been obtained over the best SoTA multimodal model while enjoying low latency and energy levels. This is attributed to the hierarchical design space that first searches for an optimal network to learn the unimodal embedding of features and then proceeds to optimize the joint embedding through the fusion architecture search. Furthermore, *Harmonic-NAS* has also shown an adaptation to different

hardware devices with varying computational capacities. The Pareto fronts in Figure 3 and the selected Pareto models in Table 4 show such hardware efficiency diversity, highlighting further the effectiveness of *Harmonic-NAS* in designing efficient MM-NNs on Edge devices.

② **The Multimodal IMDB:** In Table 5, we show the performance of the Pareto optimal backbones and multimodal models of *Harmonic-NAS* compared to SoTA methods on MM-IMDB dataset.

As a metric for the prediction performance, we use the weighted F1-score (F1-W) instead of the TOP-1 accuracy to account for the imbalanced data distribution in the MM-IMDB dataset. As shown, *Harmonic-NAS* achieves superior results, surpassing the best results achieved by SoTA works such as BM-NAS Yin et al. (2022) with up to $\sim 1.45\%$ improvement in the weighted F1-score.

On the hardware side, *Harmonic-NAS* models are $\sim 4.18x$ and $\sim 2.19x$ more latency and energy efficient compared to BM-NAS multimodal Yin et al. (2022) models. Additionally, the least accurate multimodal network in *Harmonic-NAS* is $\sim 2.89x$ and $\sim 6.14x$ more latency and energy efficient than the most accurate SoTA model on the Jetson AGX. This further demonstrates our framework’s superiority in adapting to scenarios where hardware efficiency is prioritized.

③ **The Harmful Politics Memes:** In Table 6, we report the performance of the Pareto optimal backbones and multimodal models of *Harmonic-NAS* and SoTA works on Memes-P dataset.

Compared to SoTA works, our framework provides better unimodal and multimodal neural networks regarding accuracy and hardware efficiency. The optimal multimodal networks found by *Harmonic-NAS* for both hardware devices have shown an improvement of up to $\sim 3\%$ in the TOP-1 accuracy while being $\sim 6.06x$ more latency efficient than MOMENTA Pramanick et al. (2021) on the AGX. In light of highlighting the hardware awareness of *Harmonic-NAS*, we also found another MM-NN with a slight drop in accuracy but more suitable for resource-constrained scenarios, exhibiting $\sim 1.19x$ and $\sim 1.32x$ gains in latency and energy, respectively, compared to the most accurate MM-NN from *Harmonic-NAS*.

Table 5: Performance evaluation on MM-IMDB dataset

SoTA work	Modality	F1-W(%)	Latency(ms)	Energy(mj)
Unimodal Backbones				
Maxout Yin et al. (2022)	Text	57.54	A:0.82, T:0.99	A:3.33, T:1.27
Maxout (Ours)	Text	61.18	A:0.62, T:1.09	A:2.87, T:1.40
VGG19	Image	49.21	A:30.91, T:117.94	A:791.16, T:1013.09
Harmonic-NAS (T:TX2)	Image	46.12	39.82	248.05
Harmonic-NAS (A:AGX)	Image	46.12	20.09	264.34
Multimodal Neural Networks				
MFAS Pérez-Rúa et al. (2019)	Image + Text	62.50	A:32.77, T:119.97	A:800.2, T: 1016.4
BM-NAS Yin et al. (2022)	Image + Text	62.92	A:32.78, T:119.97	A:800.66, T:1016.58
		63.61	21.37	113.99
Harmonic-NAS (TX2)	Image + Text	64.36	28.68	163.04
		64.27	23.67	121.75
		63.75	11.32	130.42
Harmonic-NAS (AGX)	Image + Text	64.36	14.29	177.55
		64.27	13.05	140.00

Table 6: Performance evaluation on Memes-P dataset

SoTA work	Modality	Acc(%)	Latency(ms)	Energy(mj)
Unimodal Backbones				
TextBERT	Text	74.55	A:22.71, T:59.63	A:524.28, T:461.91
Maxout (Ours)	Text	83.38	A:0.61, T:1.08	A:2.82, T:1.21
VGG19	Image	73.65	A:31.01, T:116.80	A:786.94, T:1001.17
DenseNet-161	Image	71.80	A:28.59, T:92.32	A:589.54, T:710.21
ResNet-152	Image	71.02	A:29.80, T:92.58	A:754.56, T:798.52
Harmonic-NAS (T:TX2)	Image	84.78	10.40	28.88
Harmonic-NAS (A:AGX)	Image	84.78	5.68	38.28
Multimodal Neural Networks				
VILBERT CC Lu et al. (2019)	Image + Text	84.66	A:22.58, T:59.93	A:523.93, T:449.89
MOMENTA Pramanick et al. (2021)	Image + Text	87.14	A:35.09, T:176.43	A:1311.22, T:1463.08
		88.45	10.51	25.63
Harmonic-NAS (TX2)	Image + Text	90.42	12.47	31.92
		90.14	11.11	26.63
		88.16	5.79	38.34
Harmonic-NAS (AGX)	Image + Text	90.42	7.51	50.79
		90.14	6.91	43.49

4.2.3. ON THE IMPORTANCE OF THE HIERARCHICAL DESIGN FOR MULTIMODAL NETWORKS

To further demonstrate the importance of the considered neural design parameters for multimodal networks in *Harmonic-NAS*, we provide Figure 4 in which we report the obtained performances when progressively adding new design dimensions to the search space. In the x-axis, we refer to the fixed and searched backbones by 'FB' and 'SB', respectively. The fixed and searchable fusion

networks are referred by 'FF' and 'SF', respectively. The term 'HW' in the x-axis designates the inclusion of the Hardware efficiency metrics (i.e., latency and energy) as optimization objectives.

In the first case ($FB+FF$), the unimodal backbones are fixed for each modality to the subnets with the highest learning capacity from our supernets (i.e., the subnets with the maximum number of

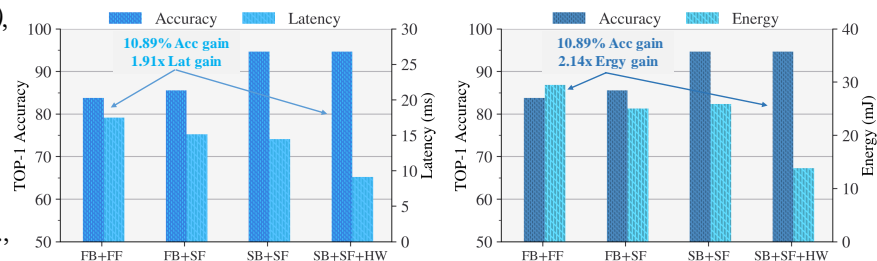


Figure 4: Performance of the progressive design of multimodal networks

learnable weights), whereas the fusion is selected as a dense architecture with all the fusion possibilities. In the second and third cases, we activate the search for the fusion network in ($FB+SF$) and unimodal backbones in ($SB+SF$). All the results in Figure 4 are reported for the AV-MNIST dataset on the NVIDIA Jetson TX2. From comparing ($FB+FF$) and ($FB+SF$), we notice that simply optimizing the fusion network for fixed backbones does not always result in optimal multimodal models. This supports our assumption of the importance of jointly optimizing the unimodal and multimodal feature embedding, as the best standalone unimodal backbones are not initially designed nor trained for multimodal learning. In ($SB+SF$), incorporating the unimodal backbones optimization yields an accuracy improvement of $\sim 11\%$. This is because our first-stage search engine could identify unimodal backbones tailored for the multimodal scenario, even if their unimodal performances are lower than that of the maximum subnets. Furthermore, by adding the Hardware metrics as optimization objectives in the ($SB+SF+HW$) case, latency and energy gains of $\sim 1.91x$ and $\sim 2.14x$, respectively, have been obtained compared to the ($FB+FF$) case while ensuring the same accuracy level, emphasizing further the importance of considering the hardware efficiency when designing multimodal networks on resource-constrained Edge devices.

5. Conclusion

In this paper, we presented *Harmonic-NAS*, a novel framework for the joint optimization of unimodal backbones and fusion networks to learn an effective joint embedding of features from multiple modalities. *Harmonic-NAS* employs a two-tier optimization scheme with an evolutionary search stage for the unimodal backbone networks and a differentiable search stage for the fusion architecture design. *Harmonic-NAS* also includes the hardware dimension in its optimization procedure by integrating metrics such as latency and energy consumption. Evaluation results have seen the superiority of *Harmonic-NAS* over SoTA multimodal-NAS approaches in discovering efficient multimodal networks with up to $\sim 10.9\%$ accuracy improvement, $\sim 1.91x$ latency reduction, and $\sim 2.14x$ energy efficiency gain on different Edge devices.

References

Halima Bouzidi, Mohanad Odema, Hamza Ouarnoughi, Mohammad Abdullah Al Faruque, and Smail Niar. Hadas: Hardware-aware dynamic neural architecture search for edge performance scaling. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6, 2023. doi: 10.23919/DATE56975.2023.10137095.

- Han Cai et al. Proxylessnas: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations (ICLR)*, 2018.
- Han Cai et al. Once-for-all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations (ICLR)*, 2019.
- Jie Cai et al. Multimodal continual graph learning with neural architecture search. In *Proceedings of the ACM Web Conference 2022*, pages 1292–1300, 2022.
- Kalyanmoy Deb et al. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *Parallel Problem Solving from Nature PPSN VI*. Springer, 2000.
- Xuanyi Dong and Yi Yang. One-shot neural architecture search via self-evaluated template network. In *ICCV*, pages 3681–3690, 2019.
- Thomas Elsken et al. Neural architecture search: A survey. *JMLR*, 20(1):1997–2017, 2019.
- Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, and Zijun Yu. Circulant-interactive transformer with dimension-aware fusion for multimodal sentiment analysis. In *ACML*, pages 391–406. PMLR, 2023.
- Xiaofeng Hou et al. Analyzing the hardware-software implications of multi-modal dnn workloads using mmbench. *arXiv:2212.01241*, 2022.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.
- Yu Huang et al. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.
- Zheng Jian et al. Eenas: An efficient evolutionary algorithm for neural architecture search. In *ACML*, pages 1261–1276. PMLR, 2023.
- Alistair EW Johnson et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1): 1–9, 2016.
- Hamid Reza Vaezi Joze et al. Mmtm: Multimodal transfer module for cnn fusion. In *CVPR*, pages 13289–13299, 2020.
- P Liang et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Conference (Neurips)*, 2021.
- Chenxi Liu et al. Progressive neural architecture search. In *ECCV*, pages 19–34, 2018a.
- Hanxiao Liu et al. DARTS: Differentiable architecture search. In *International Conference on Learning Representations (ICLR)*, 2019.
- Risheng Liu et al. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1600–1608, 2021.
- Zhun Liu et al. Efficient low-rank multimodal fusion with modality-specific factors. In *Annual Meeting of the Association for Computational Linguistics*, 2018b.

- Jiasen Lu et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Neurips*, 32, 2019.
- Pavlo Molchanov et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *CVPR*, pages 4207–4215, 2016.
- Basil Mustafa et al. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.
- Arsha Nagrani et al. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021.
- Mohanad Odema, Halima Bouzidi, Hamza Ouarnoughi, Smail Niar, and Mohammad Abdullah Al Faruque. Magnas: A mapping-aware graph neural architecture search framework for heterogeneous mpsoc deployment. *arXiv preprint arXiv:2307.08065*, 2023.
- Juan-Manuel Pérez-Rúa et al. Mfas: Multimodal fusion architecture search. In *CVPR*, 2019.
- Shraman Pramanick et al. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455. Association for Computational Linguistics, November 2021.
- Hasib-Al Rashid et al. Tinym2net-v2: A compact low power software hardware architecture for multi modal deep neural networks. *ACM Transactions on Embedded Computing Systems*, 2023.
- Luis R Soenksen et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, 5(1):149, 2022.
- Valentin Vielzeuf et al. Centralnet: a multilayer approach for multimodal fusion. In *ECCV*, 2018.
- Dilin Wang et al. Alphanet: Improved training of supernets with alpha-divergence. In *ICML*, pages 10760–10771. PMLR, 2021.
- Bichen Wu et al. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*, pages 10734–10742, 2019.
- Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.
- Zhen Xu, David R So, and Andrew M Dai. Mufasa: Multimodal fusion architecture search for electronic health records. In *AAAI*, volume 35, pages 10532–10540, 2021.
- Yihang Yin et al. Bm-nas: Bilevel multimodal neural architecture search. In *AAAI*, volume 36, pages 8901–8909, 2022.
- Jiahui Yu and Thomas S Huang. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.
- Zhou Yu et al. Deep multimodal neural architecture search. In *MM*, 2020.
- Xinyi Zhang et al. H2h: heterogeneous model to heterogeneous system mapping with computation and communication awareness. In *DAC*, 2022.