|  |  |  |
| --- | --- | --- |
| **IJIBC 25 3 - x** MRO: Multimodal Routing Optimization viaNeural Architecture Search of Fusion Paths Jeong-Hun Kim, Mi-Hwa Song\*  *Division of Computer Engeneering, Hansung University*  [*bandlagom0927@hansung.ac.kr*](mailto:bandlagom0927@hansung.ac.kr)*, mhsong@hansung.ac.kr ,*   |  | | --- | | *Abstract* *In multimodal learning, the effective fusion of interactions between different modalities is a key factor in enhancing performance, leading to improved system efficiency. However, the existing Cross-Modal Transformer (CMT)-based fusion method relies on fixed modality-specific attention paths and has limitations in flexibly exploring fusion strategies that are optimized for diverse inputs. In addition, as the computational cost increases exponentially with the number of modalities, performance bottlenecks occur and scalability across modalities becomes limited. We propose a Multimodal Routing Optimization (MRO) framework that selects the optimal routing configuration, balancing predictive accuracy and computational efficiency (FLOPs) via Neural Architecture Search (NAS). Inspired by the Once-for-All (OFA) paradigm, the MRO framework defines the attention paths between modalities as a Supernet structure.* | | ***Keywords:*** *Neural Architecture Search, Multimodal Learning, Fusion Paths, MRO* | |

**1. Introduction**

The fusion of several different types of information is central to multimodal learning. These types of information include text, voice and images. This field strives to achieve a more profound comprehension. It does this by integrating these diverse data sources. The Cross-Modal Transformer (CMT)[[1]](#one) architecture is used across many research projects to deal with complex interactions between different types of data. CMT allows for precise modeling of cross-modal semantic transitions by assigning one modality as the query and performing attention over the others.

Two main approaches to the existing CMT-based fusion method are identified. The first is a Pairwise Full Attention structure that performs cross-modal attention across all pairs of input modalities. The former offers high expressive power, but as the number of modalities (k) increases, the computational cost scales quadratically (O(k²)), leading to bottlenecks. Conversely, while the latter can reduce computation time, its fixed path design means it is unable to respond flexibly to changes in data quality or input diversity, and it is prone to performance instability.

To solve this problem, MRO treats attention paths as search variables, expands the Pairwise CMT with all cross-modal paths into a Supernet, and defines it as a searchable structure using binary masks applied to each attention path. It then selects the optimal subnet by considering the trade-off between accuracy and computational cost (FLOPs) through Neural Architecture Search (NAS).

Manuscript Received: xxxxxx / Revised: xxxxxx / Accepted: xxxxxx

Corresponding Author: mhsong@hansung.ac.kr

Tel: +82-2-760-4124, Fax: +82-2-760-5771

Associate Professor, Division of Computer Engineering, Hansung University, Korea

**2. Related Works**

**2.1 NAS: From reinforcement-based search to dynamic fusion strategies.**

The Neural Architecture Search (NAS) [[2]](#two) method comprises the following three factors. Firstly, there is the exploration of neural networks by automatic means. Secondly, there is the navigation algorithm. Thirdly, there is the performance evaluation algorithm. In the early stages, network architectures are repeatedly sampled through a reinforcement learning-based controller. Subsequently, NASNet was successfully applied to image classification. Since that seminal paper, various evolutionary algorithms have been developed and there have been significant advances in their practical application. The present study aims to convert the process of path selection to explore the path selection process itself, as opposed to the use of a fixed multidimensional fusion structure or fixed multi-month convergence structure.

**2.2 Lightweight Supernet-based NAS: Once-for-All (OFA)**

The Once-for-All (OFA)[[3]](#three) concept proposes a flexible framework to support different neural network configurations within one large-scale supernet. his approach has been proven to be efficient in that it utilizes masking techniques to enable rapid evaluation and deployment of substructure networks through masked one-shot learning without the need for retraining. Building on the core principles of OFA, this study extends the Pairwise CMT model into a supernet structure to explore effective routing path combinations suitable for various multimodal contexts. Such a design can adaptively support multimodal convergence even in situations with limited data availability, and at the same time, it is a promising way to improve computational efficiency

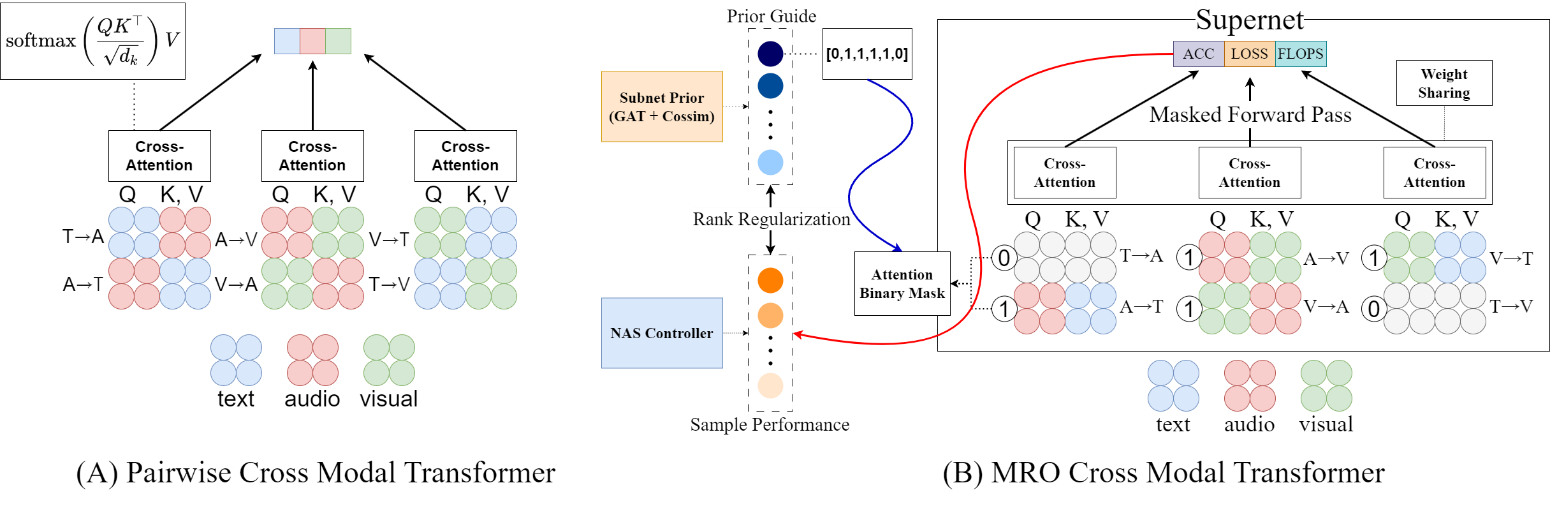
**2.3 Prior-guided Constrained Search**

The issue of expanding the navigation space has been the focus of recent research in the field of NAS. This is to facilitate effective navigation. It is also to establish appropriate search efficiency and structural properties. One notable study involves PGONAS [[4]](#four). The alignment between the subnet candidate group ranking and the subsequent performance of the subnet candidate group and the actual performance is demonstrated by this.

Meanwhile, GraphCFC [[5]](#five) formalizes modality interactions as conditionally connected graphs, using them as prior distributions to guide early NAS exploration. This approach prioritizes meaningful subnetwork configurations, improving search efficiency by avoiding random, uninformative sampling.

**3. MRO : Multimodal Routing Optimization**

**3.1 MRO Overview**



**Figure 1. MRO Structure**

As illustrated in Figure 1, the architectural differences between the baseline Pairwise Cross-Modal Transformer (CMT) and the proposed Multimodal Routing Optimization (MRO) framework are evident.

(A) illustrates the Pairwise CMT architecture, the pairwise attention scheme proposed in One-vs-Other [6]. The model performs all the cross-attention paths (T→A, T→V, A→T, A→V, V→T, and V→A) between all modality pairs, exhibiting high levels of expressive power. However, as the number of modalities increases, the amount of computation increases rapidly and a problem of lack of flexibility due to restructuring occurs.

(B) illustrates the proposed MRO framework, which expands the Pairwise CMT into a Supernet and selectively controls whether each path is activated through the Attention Binary Mask. MRO uses a GAT (Graph Attention Network)[[7]](#seven)-based directional graph and cosine similarity to measure the importance of interactions between modalities and create a Prior Guide. The purpose of the Prior Guide is to select more promising subnet configurations during the initial exploration phase of the NAS. The selected subnets are attention masked to fix their structure, and the NAS controller updates its parameters based on the performance of those masked subnets. Both the evaluation and granularity are controlled during this process.

In this process, rank regularization between prior and sample performance is applied. This is an important step in minimizing prior-based search bias and selecting subnets that match actual performance more closely. The proposed approach is designed to optimize the computational efficiency of the learning process, utilizing a masked one-shot learning method with weight sharing within Supernet.

**3.2 Supernet Architecture for Cross-Modal Attention Path Exploration**

The proposed framework integrates all pairwise attention routes among three modalities – text, audio, and vision – into a unified Cross-Modal Transformer (CMT) Supernet. Specifically, six cross-modal attention paths (T→A, T→V, A→T, A→V, V→T, V→A) are each implemented as independent attention blocks within the Supernet, and their activation is controlled via a binary mask vector.

Each constituent element of the Supernet is made up of standard Transformer components, which have been implemented in PyTorch. These components include multi-head self-attention, layer normalization and a feedforward network. Linear transformation layers are used to project each input modality into a fixed embedding space. During the forward pass, solely the attention blocks corresponding to active mask entries (i.e. mask value = 1) are activated and executed, while inactive paths are wholly bypassed to reduce computational overhead. Consider, for example, a mask vector of [1, 0, 1, 1, 0, 1]. This activates four out of six possible attention paths, and it is only those that contribute to the fusion process. The outputs from the active paths are aggregated via average pooling and passed to a shared classification head for final prediction. A weight-sharing paradigm is used to train the Supernet to facilitate efficient neural architecture search (NAS). In this paradigm, all possible subnets reuse the same parameters, meaning retraining is not necessary. This facilitates rapid evaluation of multiple subnet configurations during the NAS process. It does this by significantly reducing computational costs. At the same time, it maintains structural flexibility.

**Table 1. Pseudocode for NAS-Based Exploration of Cross-Modal Transformer Paths**

|  |
| --- |
| Supernet ← build\_CMT\_Supernet()  Prior ← compute\_initial\_prior()  for epoch in range(E):  if epoch < Warmup:  Masks ← uniform\_sample(n)  else:  Masks ← prior\_guided\_sample(Prior, n)  Losses ← []  for mask in Masks:  Subnet ← Supernet(mask)  L ← TaskLoss(Subnet)  optimize(L)  Losses.append(L)  R ← 0  for (i, j) in pairwise(Masks):  R += max(0, (Losses[i] - Losses[j]) \* sign(Prior[Masks[i]] - Prior[Masks[j]]))  optimize(R)  Prior ← update\_prior(Masks, Losses)  TopK ← select\_final\_subnets(Prior, Logs) |

Table 1 illustrates the Neural Architecture Search (NAS) process. This process identifies the optimal configuration of attention paths. Initially, a prior distribution over all attention pathways is computed based on semantic similarity scores derived from the Modal Relation Graph. Binary attention masks are sampled uniformly during the warm-up round. This ensures diverse exploration of the search space. Once the warm-up round has ended, the algorithm switches to a prior-guided sampling strategy. In this strategy, the sampling of subnet candidates is with probabilities that are biased by the learned prior. Each sampled mask activates a specific subset of attention paths within the supernet. A distinct subnet architecture is defined by this. For each subnet, the task-specific loss is subject to calculation. This could be cross-entropy for classification. The loss is then backpropagated, causing an update to the shared weights in the supernet. A rank regularization term is introduced to align the prior expectation and the observed subnet performance. This is achieved by promoting consistency in relative ranking. Mismatches between the predicted ranking and the observed ranking are penalized by this term. The predicted ranking is from the prior. The observed ranking is based on loss values. The regularization loss is added to the overall objective and then optimized as part of the overall process. After each epoch, the prior distribution is updated based on the performance of the sampled subnets. This iterative update process helps guide the search towards promising regions of the architecture space. The best subnets for final review or use are selected once the search is complete. The selection is based on the highest prior scores and/or the best validation performance, with these factors given the greatest weight. The proposed MRO framework is made scalable and practical for real-world multimodal applications by this masked one-shot NAS strategy, which allows it to dynamically identify efficient and high-performing attention configurations without retraining or exhaustive search.

**3.3 Prior-Guided Search Using a Modal Relation Graph**

In this study, the Model Relation Graph (MRG) is proposed to represent the semantic structure of the navigation space and to serve as a guiding model in the initial stage of architecture research. The MRG is defined as a directed graph where each node represents a modality, and edges indicate the direction and strength of interaction based on semantic similarity and interaction relevance.

(1)

Specifically, the edge weights are calculated by calculating three indicators for any modal pair i and j, with the first being the semantic similarity between the two modalities, Proximity in the representation space is measured using the cosine similarity between the embedding vectors of i and j. Second, the gradient variance that occurred in the corresponding path during the learning process indicates how much the path contributed to the update of model parameters, and we estimate the importance based on the amount of change in the gradient signal. Third, the attention weight variance in the corresponding modal path quantifies the diversity and dispersion of information flowing through the path and evaluates the stability and expressive power distribution of interactions between modalities. The edge importance between modal i and modal j is calculated based on three metrics. First, we set weights α, β, and γ corresponding to each metric, and then multiply them by the values of cosine similarity (i, j), gradient variance (i, j), and activation variance (i, j), respectively, and sum them to compute a composite score for edge e(i→j). This approach provides a structural measure that simultaneously reflects semantic similarity between modalities and sensitivity (gradient/activation variability) during the learning process. The resulting edge scores are fed into a Graph Attention Network (GAT), which coordinates interactions between neighboring modalities in an attention-aware manner. GAT learns the weights that each modality i assigns to other modalities j to which it is connected, and these values are transformed into a probabilistic prior distribution via SoftMax normalization.

This prior distribution is used to define the probability of selecting each attention path during the Neural Architecture Search (NAS) exploration process. As a result, semantically important paths with high interaction strengths are preferentially selected, which increases the efficiency of structure discovery and contributes to improved model representation and generalization performance.

**3.4 Prior-Guided Search with Rank Regularization for Consistency Alignment**

Our approach utilizes priority-guided NAS exploration to prioritize important attention paths during the exploration process by leveraging semantically informed prior knowledge extracted from modal relation graphs (MRGs). To address potential discrepancies between semantic relevance and actual performance, we introduce a ranking normalization loss that aligns prior-based and performance-based rankings. This mechanism helps balance semantic guidance with empirical results progressively and gradually, leading to more stable, goal-oriented convergence. The formal definition of the ranking normalization loss is as follows:

(2)

 In this case, the total is made up of the individual ranking and the performance-based ranking. As an alternative option, it can maximize the correlation coefficient of ranking consistency. This dynamic alignment method gradually improves the reliability of the prior distribution throughout the navigation process, and the NAS controller reasonably prefers excellent subnet structures. As navigation progresses, the architecture space is gradually refined, and ultimately, the optimal subnet configuration is achieved by balancing predictive accuracy and computational efficiency.

**4. Evaluation**

**4.1 Learning Dataset**

**Table 2. Distribution of 7-Class Emotion Labels in CMU-MOSEI and MELD**

|  |  |  |
| --- | --- | --- |
| Dataset | CMU\_MOSEI | MELD |
| sentiment | Mean Probability | Mean Probability |
| Happy (Joy) | 27.76% | 15.8% |
| Sadness | 32.23% | 9.0% |
| Angry | 14.61% | 14.3% |
| Disgust | 10.51% | 3.0% |
| Fear | 3.4% | 2.5% |
| Surprise | 6.49% | 11.3% |
| Neutral | 5.0% | 44.1% |
| total | 100% | 100% |

This study independently used CMU-MOSEI [[8]](#eight) and MELD [[9]](#nine), public multimodal emotion recognition datasets, in the experiment to verify the performance and lightweight effect of the proposed Multimodal Routing Optimization (MRO) structure from various angles. The two datasets complement each other in terms of their composition, speech units, modal combinations and labelling systems, making them suitable for evaluating the performance of MRO under various input conditions. CMU-MOSEI consists of approximately 23,500 single speaker utterances, with each sample containing text, voice, and facial expression, and is labelled with seven emotion classes. The structure is based on a single speaker, and the time consistency and interaction between modalities are clear. This provides a basic structure verification environment for MRO to maintain performance while reducing the amount of computation. MELD is a multimodal dataset consisting of over 13,000 utterances from the TV show Friends, aligned across text, speech, and visual modalities. It records conversations between speakers, how they flow, and how emotions change. This makes it a good way to measure choosing paths and understanding the situation, which are important parts of MRO. Distinct modality characteristics are possessed by CMU-MOSEI and MELD, so Attention Binary Mask optimization is performed separately, and the two datasets are trained independently without parameter sharing. This allows for a systematic validation of MRO's robustness and adaptability.

**4.2 Model Hyperparameter**

**Table 3. Hyperparameter Settings for MRO on the CMU-MOSEI and MELD Datasets**

|  |  |  |
| --- | --- | --- |
| Dataset | CMU\_MOSEI | MELD |
| Batch size | 32 | 64 |
| epoch | 40 | 30 |
| Learning rate | 3e-5 | 5e-5 |
| Dropout | 0.4 | 0.3 |
| Model dim | 256 | 256 |

The model hyperparameters were set to consider both learning stability and convergence speed, considering the number of samples, utterance length and emotional distribution characteristics of each dataset. CMU-MOSEI and MELD have similar modal configurations, but since they show differences in data size and context complexity, different optimal settings are required even under the same structure. In the case of CMU-MOSEI, there are more than 23,000 samples, utterances are relatively long, and modalities are highly consistent. Taking this into account, the batch size was set to 32 and learning was performed over a relatively long period of 40 epochs. Additionally, the learning rate was set to 3e-5 to achieve stable convergence and a dropout rate of 0.4 was applied to prevent overfitting. The multimodal embedding dimension was set to 256 to ensure consistency of embedding dimensions across modalities in MRO. MELD is a dataset that utilizes a variety of communication modalities, characterized by rapid topic switching and frequent short sentences. The total number of samples is about 13,000, which is smaller than MOSEI. Therefore, the batch size was set to 64 and the training epoch was adjusted to 30 to improve learning efficiency and reduce the risk of overfitting. In addition, the learning rate was set to 5e-5 to drive fast convergence, and the dropout rate was set to 0.3 to ensure good generalization performance.

**4.3 Performance Results**

Four multimodal fusion structures were included in the performance comparison experiments. First, the Concat + MLP method concatenates the outputs of each modality and classifies them using a multilayer perceptron (MLP). Second, EmbraceNet is a structure designed to maintain robust performance even when some modalities are missing. Third, the Cross-Modal Transformer (CMT) performs attention-based fusion across all modality pairs, exhibiting high expressive power but incurring high computational cost. Finally, the MRO structure proposed in this study is based on the CMT.

**Table 4. Performance and FLOPs of MRO and Baselines on CMU-MOSEI and MELD**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Model | MAE | Corr | Acc-7 | Acc-2 | F1 | MFLOPs |
| CMU\_MOSEI | Concat+MLP | 0.645 | 0.601 | 41.20 | 77.85 | 77.40 | 22.31 |
| EmbraceNet | 0.602 | 0.644 | 44.30 | 79.10 | 78.95 | 26.78 |
| Pairwise CMT | 0.558 | 0.721 | 50.90 | 84.20 | 83.80 | 254.94 |
| MRO(Mask[0,1,1,0,1,1]) | 0.519 | 0.763 | 55.30 | 86.45 | 86.30 | 173.48 |
| MELD | Concat+MLP | 0.634 | 0.589 | 61.83 | 85.67 | 83.12 | 23.90 |
| EmbraceNet | 0.598 | 0.631 | 63.92 | 87.04 | 84.55 | 29.20 |
| Pairwise CMT | 0.561 | 0.683 | 66.14 | 89.36 | 86.40 | 378.82 |
| MRO(Mask[0,1,1,1,1,0]) | 0.559 | 0.691 | 67.07 | 90.32 | 88.03 | 297.36 |

Table 3 shows the results of comparing the performance indicators (MAE, Corr, Acc-7, Acc-2 and F1) and the amount of computation (MFLOPs) required by each model for the CMU-MOSEI and MELD datasets. The experiment revealed that MRO outperformed Pairwise CMT in both datasets, while significantly reducing the amount of computation by utilizing only certain attention paths. For instance, on the CMU-MOSEI dataset, MRO achieved an Acc-7 score of 55.30%, representing a 4.4% improvement, while the computation required decreased by approximately 32% to 173.48 MFLOPs. A similar trend was observed in the MELD dataset: Acc-7 improved by 0.93% (66.14% to 67.07%), and the amount of computation decreased by 21%.  
 In particular, the difference in accuracy between the two datasets is due to variations in the shapes of the labels. Since the CMU-MOSEI dataset provides emotion distributions in the form of soft labels with continuous probabilities (e.g. neutral: 0.6; sadness: 0.3; etc.), there are many samples containing mixed emotions with unclear boundaries. Consequently, the model is trained to predict the distribution between multiple emotions, resulting in relatively low accuracy, as a clear match criterion with the correct emotion (class) is required. Conversely, MELD consists of clear, single, hard-label-based emotion classifications, achieving higher overall accuracy in the Acc-7 and Acc-2 indicators. In contrast, CMU-MOSEI tends to record higher F1 scores than MELD (CMU 86.30, MELD 66.88 based on MRO). This is because CMU-MOSEI learnt the predictive balance between classes through soft label-based emotion distribution prediction and demonstrated more robust performance in terms of precision and reproducibility when considering the similarities and boundaries between emotions.

F1 scores are an important metric for evaluating model performance in situations where there is imbalance between classes or where the boundaries between emotions are ambiguous. The CMU-MOSEI dataset used in this study often contains multiple emotions simultaneously within a single utterance or subtle emotional differences, resulting in prominent emotion overlaps, which increases the difficulty of classification. On the other hand, MELD has clearly distinguished emotion labels, making classification criteria explicit; however, it suffers from severe class imbalance in data distribution, leading to prediction bias toward specific emotions and resulting in relatively low F1 scores. In this context, the proposed MRO routing mechanism effectively addresses the uncertainty of soft label-based prediction and maintains stable convergence performance across various multimodal input environments.

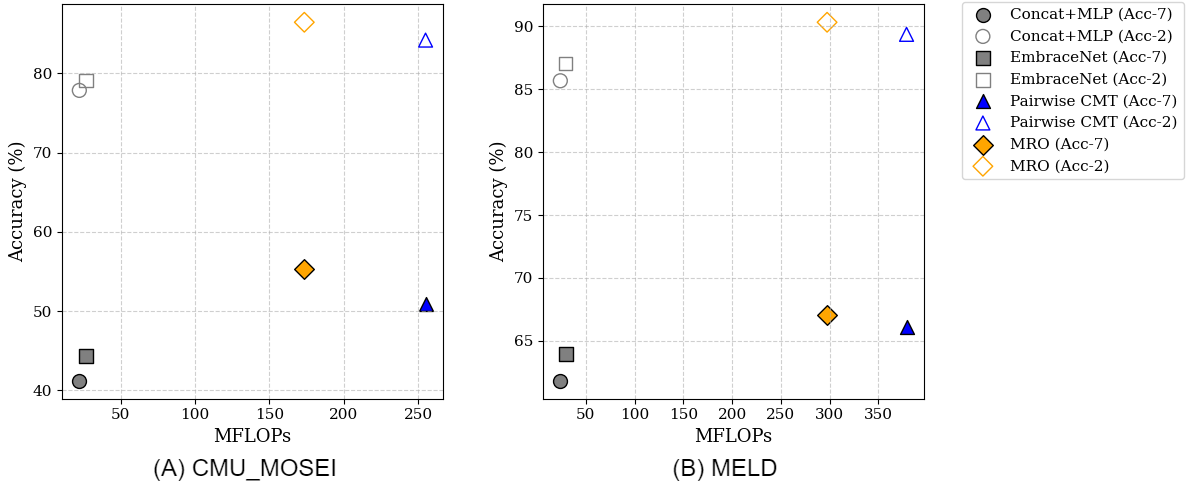


Figure 2. Accuracy – FLOPs Tradeoff

Figure 2 shows a trade-off curve visualizing the accuracy of each multimodal fusion model (Acc-7 and Acc-2) according to the amount of computation (MFLOPs) based on the CMU-MOSEI (left) and MELD (right) datasets. The CMU-MOSEI-based experimental results presented on the left show that basic structures such as Concat+MLP and EmbraceNet perform relatively poorly in terms of prediction accuracy, despite the advantage of low computation. The experimental results show that the Pairwise CMT model, which comprehensively utilizes cross-modal attention paths between all modalities, achieved high accuracy but has the drawback of very high computational complexity.

MRO achieves better performance and computational efficiency than traditional CMT models. It achieves slightly higher accuracy than CMT on both the Acc-7 and Acc-2 criteria with fewer concurrent computations (MFLOPs). Specifically, on the MELD dataset, MRO saves approximately 21% of FLOPs, despite the fact that the emotion classification accuracy (Acc-2) is almost identical. These results demonstrate that, rather than using all paths, MRO efficiently utilizes computational resources by selectively activating only meaningful attention paths. It remains expressive while eliminating unnecessary paths, thus increasing efficiency without compromising performance.

As can be seen in Figure 2, MRO is located at the top left of the accuracy-computation balance, providing visual confirmation of Pareto efficiency. Beyond numerical comparisons, this helps us to intuitively understand the trade-off between performance and efficiency in different structures.

**5. Conclusions**

The results of experimenting with CMT in pairwise structures via the MRO framework are valid. Experiments using MELD and CMU\_MOSEI datasets show that MRO maintains or improves accuracy. This occurs despite being trained by activating fewer paths than CMT. This is based on the number of attention paths (masks) used. We show that dynamic Attention Path selection based on relationships and similarities between modalities reduced both interactions and weights.

In the future, the MRO framework will not be limited to traditional multimedia modalities such as text, voice and images. Instead, it will be able to expand effectively in terms of structural flexibility and resource efficiency, even in high-quality, high-dimensional multimodal environments such as medical images (MRI and CT), Internet of Things (IoT) sensors, predictive preservation (PHM) and satellite images. The MRO will be applied to these practical domains in future to quantitatively verify the effect on performance and computational efficiency in the actual system environment.

**References**

[1] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal Transformer for Unaligned Multimodal Language Sequences,” *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019. DOI: <https://doi.org/10.18653/v1/p19-1656>

[2] B. Zoph and Q. V. Le, “Neural Architecture Search with Reinforcement Learning,” *arXiv preprint arXiv:1611.01578*, 2016. DOI: <https://doi.org/10.48550/arXiv.1611.01578>

[3] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, “Once-for-All: Train One Network and Specialize it for Efficient Deployment,” *Proc. Int. Conf. on Learning Representations (ICLR)*, 2020.

DOI: <https://doi.org/10.48550/arXiv.1908.09791>

[4] P. Dong, X. Niu, L. Li, L. Xie, W. Zou, T. Ye, Z. Wei, and H. Pan, “Prior-Guided One-shot Neural Architecture Search,” *Proc. AAAI Conf. on Artificial Intelligence*, vol. 38, no. 7, pp. 8947–8955, 2024. DOI: <https://doi.org/10.48550/arXiv.2206.13329>

[5] J. Wang, X. Wu, Y. Song, and S. Liu, “GraphCFC: An Efficient Graph-Based Method for Conditional Fusion Control,” *IEEE Trans. Multimedia (TMM)*, 2023. DOI: <https://doi.org/10.1109/TMM.2023.3260635>

[6] M. Golovanevsky, E. Schiller, A. Nair, E. Han, R. Singh, and C. Eickhoff, “One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data,” *Proc. Int. Conf. on Artificial Intelligence in Medicine (AIME)*, 2024. DOI: <https://doi.org/10.1142/9789819807024_0041>

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, “Graph Attention Networks,” *International Conference on Learning Representations (ICLR)*, 2018.

DOI: <https://doi.org/10.48550/arXiv.1710.10903>

[8] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph,” *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. DOI: <https://doi.org/10.18653/v1/P18-1208>

[9] S. Hazarika, G. Zimmermann, S. Poria, and R. Zimmermann, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations,” *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019. DOI: <https://doi.org/10.48550/arXiv.1810.02508>